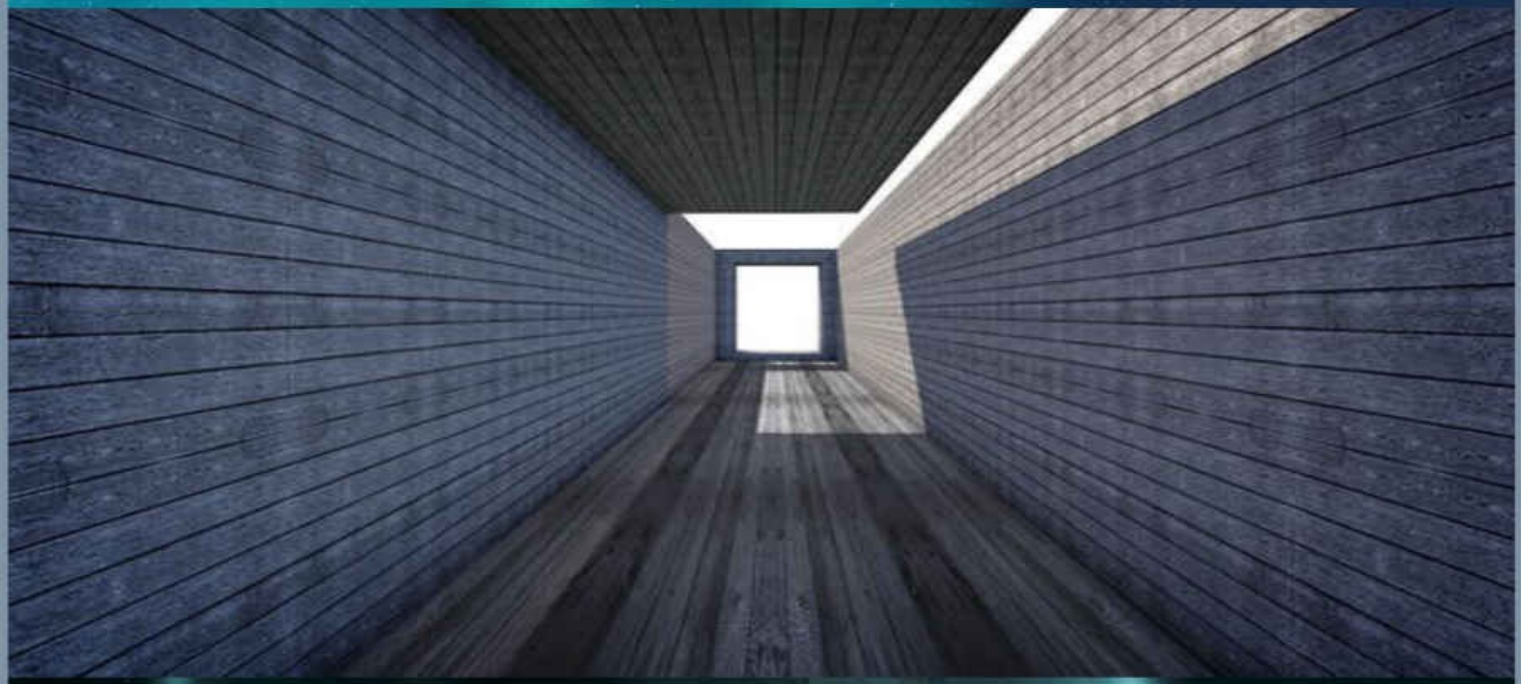
MIND, MEANING AND SCIENTIFIC EXPLANATION

BY J.-M. KUCZYNSKI



Mind, Meaning, and Scientific Explanation

by

J.-M. Kuczynski, PhD

Part I Analytic Philosophy as the Analysis of Analysis
Introduction: Philosophy as the Science of Science
Chapter 1 Analytic Philosophy as Logical Analysis
Chapter 2 Properties and Non-Spatiotemporal Existence
Part II Language
Chapter 3 What Is a Language?
Chapter 4 What Are Propositions?
Chapter 5 Definite Descriptions and the Nature of Reference
Chapter 6 Slinging the Slingshot: How a famous argument of Frege's proves his conception of reference to be incoherent
Chapter 7 Some Puzzles Relating to Language and Logic
Chapter 8 Do We Think in Words?
Chapter 9 Modality and Non-Existence
Part III Knowledge and Inductive Inference
Chapter 10 Knowledge
Chapter 11 Skepticism and the Justification of Inductive Inference
Chapter 12 Empiricism and Its Limits
Part IV Determinism, Causality, and Probability
Chapter 13 Determinism, Randomness, and Unpredictability
Chapter 14 Causality
Chapter 15 Hume's First Analysis of Causality
Chapter 16 Hume's Second Analysis of Causality

Chapter 17 Counterfactuals
Chapter 18 Explanatory vs. Statistical Probability
Chapter 19 The Raven Paradox
Part V Explanation
Chapter 20 Scientific Explanation
Chapter 21 Psychological Explanation and Interdisciplinary Reduction
Part VI Mind, Body, Self
Chapter 22 Different Views as to the Relation Between Mind and Body
Chapter 23 The Computational Theory of Mind: Progressing Towards a Scientific Psychology or Regressing to Pre-Scientific Animism?
Chapter 24 Personal and Objectual Identity
Chapter 25 The Compatibilist Conception of Freedom
Chapter 26 Selves vs. Minds, Values vs. Desires.
Chapter 27 How Is Rationalization Possible?
Chapter 28 Psychopathy, Sociopathy, and Non-Psychopathic, Non-Sociopathic Criminality
Appendix: Definitions of key terms

Part I

The Analysis of Analysis

Introduction: Philosophy as the Science of Science

1.0 Philosophy as providing the justification for the empirically unjustifiable assumptions on which empirical science depends

An *empirical* discipline is one whose findings, when justified, are justified by their consistency with *sensory observation*. Biology is an empirical discipline; so is physics.

Any given empirical discipline makes assumptions that cannot be justified *by* that discipline or, indeed, by *any* empirical discipline. For example, when investigating the external world, and thus when doing physics or biology, one must assume that one's senses give one accurate information about it. For if our sensory experiences turned out to be hallucinations, then any hypotheses based thereupon would be spurious.

But this assumption cannot *itself* be given an empirical justification. More plainly, observation cannot tell you that it itself is a source of accurate information. Witnesses cannot self-corroborate. Thus, a physicist could not coherently attempt to provide experimental support for his belief that his senses are accurate. For it is only *through* his senses that he can know whether those experiments were even conducted, as opposed to merely dreamt, or what their outcomes were, supposing that they were in fact conducted.

Of course, *to the extent* that we are able to determine to what degree, if any, this or that particular sensory experience is veridical, we do so on the basis of other sensory experiences. But when judging that, last night at 4:00 am (sleep-time), I wasn't really in the presence of a talking cow, I am assuming the veridicality of the experiences with which my cow-experiences fail to cohere. In general our sensory experiences can provide only a *relative* justification for the judgments we make about this or that specific perception: *relative* to the assumption that those experiences are typically accurate, inferences can be made as to the degree of accuracy of this or that particular perception. But we cannot legitimately assume that our sense perceptions are accurate when trying to show that sense-perception is a legitimate source of information about the external world**.** So, since the hypotheses generated by physics are based on sensory experiences, and therefore presuppose that such experiences are typically accurate, physics cannot possibly justify that presupposition, and neither, for the very same reason, can any other empirical

discipline. Therefore, unless that presupposition can be justified, we know nothing about the external world, and, as a result, all of the empirical sciences crumble. It is the task, not of the scientist, but of the *meta*-scientist---of the philosopher, in other words---to provide that justification.

The assumption that our senses are a viable source of information about the external world is but one of many assumptions on which empirical knowledge rests but for which no empirical justification can be provided and for which, therefore, a philosophical justification must be provided. It is the purpose of this book to identify the most important of these assumptions and to provide the requisite philosophical justifications for them.

2.0 The main confusion underlying empiricism

Philosophy is regarded with more suspicion than most other disciplines. One reason for this is that it is a non-empirical discipline. But that is not the only reason. Mathematics is a non-empirical discipline. But it isn't regarded with suspicion. Why this double standard?

First of all, because of the close connections of mathematics to empirical disciplines such as physics, people often assume that, like physics, mathematics must be empirical. Also, the importance of some branches of mathematics to scientific and technological innovation makes it clear that, even if mathematics *is* non-empirical, it's unwise to question its methods. Don't look a gift-horse in the mouth.

But there is another reason why, despite it's being quite as non-empirical as philosophy, mathematics is erroneously seen as empirical and why, as a result, it isn't the object of the same overt hostility and suspicion as philosophy, which, unlike mathematics, *is* seen for the non-empirical discipline that it is.

For obvious reasons, people tend to think that it is through visual observation of figures on chalk-boards (or some such) that one learns geometry and that, given any branch of mathematics, the same thing *mutatis mutandis* is true of the way in which one learns the truths belonging to it.

This reasoning embodies a confusion that has stymied intellectual progress for centuries. The message borne by a perception, or any other mental representation, is known as its "content." One way in which sense-perceptions can add to what we know is by transmitting their contents to us. I see a dog running towards me, and I know on that basis that, indeed, there is a dog running towards me. Here what I learned from my sense perception coincided with the content of that very perception.

But there is a very different way in which sense perception can add, or at least help to add, to what one knows. A sense perception may *trigger* ratiocinative processes that yield knowledge of truths that were not part of that perception's content. Many truths learned in this way couldn't *possibly* be any part of any perception's content. An illustration is in order.

You see figures on a blackboard that vaguely resemble circles, triangles, etc. These sense perceptions help you learn that:

(BH) The area of Euclidean triangle is ½ *BH*.

Those visual perceptions directly apprise you of various *non*-mathematical facts (e.g., that there is a blackboard nearby on which there are various markings). Those perceptions do not themselves *transmit* BH to you. So how do they help you become aware of that principle? The answer: By precipitating thought-processes that in their turn lead to your seeing that BH is correct.

You can see specific markings on specific surfaces; you can see specific rocks and specific people; you can see specific objects. (Given that there are no *non*-specific objects, "specific object" is a pleonasm, and we might as well have just said: "you can see *objects.*") But you cannot see (or otherwise sense-perceive) *principles*. (The parenthetical qualification will henceforth be left unstated.) There are two reasons for this:

> Reason #1: Principles are *general.* BH says that, for any *x*, *x*'s being a triangle has the consequence that *x*'s area is ½ its base times its height. One cannot *see* general truths. One sees particulars: particular rocks, trees, people, events, and so forth. One can see instances of general truths, but not general truths *per se.*

> Reason #2: Principles express dependence-relations. In other words, they are to the effect that, given *P*, *Q* must be the case, or at least is disposed to be the case, this being equivalent to: *P*'s being the case depends on (or is disposed to depend on) *Q*'s being the case. BH says that *x*'s being a triangle depends on *x*'s being such that its area is ½ its base times its height.
>
> So far as one's senses apprise one of anything, they apprise one of what is the case, not of what *must* be the case. There are several reasons for this. One of them is that *P*'s *having* to be the case is equivalent to its being impossible that *P* should fail to be the case; and one cannot sense-perceive the non-existent, let along the impossible.

Thus, our senses have two quite distinct roles in the acquisition of knowledge: an *information-transmitting* role and a *thought-triggering* role.

Here is one way of making it clear how these two roles differ. You have a

visual perception that is experientially just like that of seeing of a dog running towards you; and on that basis you believe that, indeed, there is a dog running towards you. But it turns out that you weren't really seeing a dog; you were having a hallucination. Question: did your sensory experience give you *knowledge* to the effect that there was a dog running towards you? No. It may have been that there *was* a dog running towards you. But your sense perceptions didn't provide you with knowledge of any such fact.

Bearing these points in mind, consider the following, very different scenario. You have visual experiences that are experientially just like the visual experiences you'd ordinarily have of drawings of figures on a chalkboard. Those experiences precipitate a long train of mathematically impeccable reasoning that results in your accepting BH. It turns out that your visual perceptions of the chalkboard were hallucinatory. Does it follow that you don't know that a triangle's area is ½ its base times its height?

No. Those visual experiences merely *prompted* the thought-process that led to your accepting BH. Since that thought-process was itself impeccable, even though it was initiated by hallucinations, the result was indeed knowledge of BH. Those hallucinations no more *constituted* your justification for believing BH than your turning the key in the ignition of your car *constitutes* the road-trip that you subsequently make.

But in the first scenario, any justification you had for believing that there was a dog running towards you vanished when it turned out that the experiences on which your belief was based were hallucinatory. So far as you had any justification for believing that a dog was running towards you, it lay in your visual experiences. So, given that those experiences were hallucinations, it follows that you had *no* justification for believing that a dog was running towards you and, therefore, that you didn't know it. In the other scenario, your visual experiences weren't themselves constitutive of your justification for accepting BH. They were merely precursors to the thought-processes that *did* provide you with such a justification. So the fact that those visual experiences were hallucinations in no way undermines your justification for accepting BH.

A story will help put these points into focus. A coconut lands on Newton's head. This causes him to conduct various investigations, which bear rich mathematical and scientific fruit. The blow to Newton's head wasn't constitutive of Newton's subsequent ratiocinative activity, which was impeccable; it merely initiated it. It is along these lines that one must

understand the relationship between, on the one hand your grasp of the fact that a Euclidean triangle's area is ½ its base times its height, and on the other hand the visual perceptions (of chalk figures, or whatnot) that, in due course, led to your grasping this fact.

One last scenario will make it clear just how much the reason-triggering role of sense perception differs from its information-transmitting role. Smith reads a proof of some mathematical theorem. Smith understands what he is reading and duly becomes aware of the proof in question. It is, of course, through *visual* perceptions that he becomes aware of that visual theorem. Brown hears a lecture in which that same proof is set forth. That lecture is simply a reading of the article that Smith read. (So the words Brown hears are the very words that Smith read.) Brown understands what he is hearing and duly becomes aware of the proof in question.

Smith's sense-perceptions bear no resemblance, even of the most abstract kind, to Brown's perceptions. And yet Smith and Brown come away from their respective experiences with the same mathematical knowledge. There is no way to reconcile this last fact with the supposition that it is through sense-perception that one becomes aware of mathematical truths; and it must be supposed that, in the situation just described, the visual and auditory experiences in question merely *occasioned* ratiocinative processes that, unlike those sense-perceptions, endowed in whom they occurred knowledge of the relevant theorem. Given *any* mathematical principle---or, indeed, any logical or philosophical principle----an exact analogue shows that, so far as sense-perception had a hand in a given person's acquiring knowledge of that principle it was merely by functioning as a trigger for ratiocinations that, unlike any sense-perception, were capable of awakening the relevant ratiocinative faculties.

The fact that the senses sometimes function in this way undermines the facile the claim that all knowledge is empirical and therefore delegitimizes any conceivable attempt to delegitimize the discipline of philosophy.

Let us end by making a point that will make it clear what is meant by the (already defined, but, even when properly defined, oft-misunderstood) word "empirical." When our senses *transmit* knowledge to us, the resulting knowledge is empirical. (Also, inferential and theoretical knowledge *based* on knowledge thus transmitted to us is empirical. So our knowledge of the truths described by Relativity Theory, though obviously not acquired directly from sense-perception, is empirical.) Knowledge is *not* empirical if it results

from purely ratiocinative processes, regardless of whether those processes were occasioned by sense-perceptions. Non-empirical knowledge is purely conceptual: it is knowledge, not of the structures or interrelations of denizens of the spatiotemporal world, but of the structures and interrelations of concepts. Non-empirical knowledge is typically (though probably not always) obtained by *analyzing* concepts. For this reason, it is often described as ***analytic***.

3.0 These points in relation to some facts about this book

This makes it clear that the hypotheses belonging to any given empirical discipline---be it history, botany, or astronomy---are not justified, and therefore don't constitute knowledge (even if they happen to be true) *unless* cogent justifications can be provided for assumptions that are inherently incapable of being justified on empirical grounds. Philosophy is the discipline that provides these grounds. Philosophy has completed much of its work, and there is no reason to believe, and every reason to *dis*believe, that it cannot do everything that it is its job to do.

In the present work, we'll discuss the concepts in terms of which we understand language, knowledge, thought, mind, and self. We'll see how quickly philosophical methodology, if adroitly used, can shed light on issues which, among professionals no less than the laity, are dealt with only in a prejudice-driven, sterile manner.

Some aspects of this work will be uncontroversial. Others (e.g., the discussion concerning the nature of selfhood and concerning the nature of self-deception and mind-internal conflict) are conjectural.

Philosophy is in the business of questioning foundational concepts. This means that philosophy often questions itself. This has several consequences that are of interest to those who are in the process of deciding whether or not to read this book. First, it means that, although there is a core of non-trivial statements on which analytic philosophers agree, there are foundational issues on which they don’t agree.

This means that there cannot be "textbooks" in philosophy the way there can be in biology or chemistry. A textbook is, by definition, a catalogue of frozen results. Since, in philosophy, *almost* everything is up for grabs (the "almost" is extremely important), any work purporting to be a philosophy textbook will either be trivial and not worth reading *or* it will offend practically everyone in the philosophical community. A philosophy textbook must attempt to answer many questions. Some such questions have agreed-upon answers. Most do not. This isn't because there is *no* correct answer. Given any given, be it philosophical or non-philosophical, either

(a) There is exactly one correct answer to it, and what people believe does not in any way determine what that answer is

or

(b) There is a problem with that question (e.g. it involves a false presupposition) that makes that question unanswerable, and what people believe does not in any way determine what that problem is or whether there is such a problem.

The vast majority of philosophy text book writers have chosen *not* to offend their colleagues. The result: their text books are empty and uninformative. They go over well with people who want to avoid the rigors of this discipline, which is most of the students in any given introductory philosophy class, and thus tend to make their authors popular with their students. But they're not worth reading.

3.1 Not all introductions to philosophy are empty and uninformative

There are some exceptions to this. In particular, Arthur Pap wrote two brilliant books, *Introduction to Analytic Philosophy* and *Introduction to the Philosophy of Science.* L. Susan Stebbing, and Ernest, Nagel and Morris Cohen, working jointly, wrote very good introductory texts.

The problem with these texts is that they're not quite introductory enough. They do not define technical terms; they assume that the reader will quickly master arguments that most students will not understand. (Stebbing, Pap, and Nagel, taught at Oxford, Yale, and Columbia, respectively. These are elite institutions, whose students are typically better prepared than their counterparts at less distinguished universities.) Having taught several thousand students at non-Ivy League institutions, I can confidently say that most neophytes, and many non-neophytes, will *never* master, no matter how long or hard they try, many of the concepts a mastery of which is a prerequisite for understanding any of the three books just mentioned. And, while I think that, at first-institutions, students of philosophy would, in general, be more likely to discover philosophical truths than their counterparts at bottom-tier institutions, they would not be significantly more likely to do so. This claim may seem deeply implausible. But, in the next section, I will put forth some arguments that, I believe, show it to be worthy of at least some credence. Superb though these books are as treatises, they expect more of their readers than students[6] are able to give.

The problems relate, not just to content, but to presentation. The sections into which they're divided are much too long. Neophytes need clear signposts: clear headings, clear thesis-statements, clear examples, and vivid illustrations.[7] Whenever possible, joints in the argument should be marked with abrupt changes in the physical appearance of what they’re reading. A book of this nature should be as much of a painting as a philosophical treatise should be. A 20 page section consisting of long paragraphs won't go over well with anyone, especially not neophytes, especially if the contentions being advocated aren't vividly illustrated. But rare is the philosopher who illustrates his points with examples and rarer still is the one who illustrates his points with examples whose meanings are as obscure as the statements they're intended to clarify. One reason why philosophers are so unlikely to illustrate what they're saying is that they tend not to be very creative people.

But the main reason is that, by illustrating his assertions with forceful examples, a philosopher makes it clear just what he's saying, thereby forfeiting the option of saying that critics misunderstood his work, which is more likely than not his only way of staving off devastating objections to it.

I try to write as starkly and directly as possible. I thus try to keep my prose as unclogged as possible: no disingenuous disclaimers; no gratuitous references to scholars who aren't moving the ball forward but, because they're still alive, can be a nuisance if not given what they claim to be their due. Most works in this field are replete with oily disclaimers (e.g., "granting that all of the various participants in the debate surrounding this hotly contested issue have put forth plausible and powerfully argued views, it is my tentative judgment that the preponderance of the data suggests that, quite possibly,…") that do nothing but assure the author's colleagues that any substantial point that manages to wedge its way into his prose will be rendered innocuous, buried as it is underneath the rubble of politically-driven puffery.

Because I don't write in this way, and because readers probably aren't accustomed to seeing such directness except in manifestos, many a reader will think, at least initially, that I'm being dogmatic. Such a judgment would be in error. By writing this book in a non-political fashion, I have opened up a huge amount of space that would otherwise consist of gratuitous scholarship and wormy disclaimers and, although I use much of that space to put forth claims, I use far more of it to defend them. And by "defend them," I mean *actually* defend them; I don't mean "cite several (not so) 'greats' who have advocated them"---which wouldn't really be defending them at all.

3.2 Alethic nihilism self-defeating

Alethic nihilism: The doctrine that no doctrine is true.

I've heard it said that, in philosophy, "there is no right or wrong." First of all, this is itself a philosophical statement. So if it's true, then it's false, and if it's false, then it's true. Therefore it's false, given that no true statement is ***inconsistent*** with itself. (P is inconsistent with Q if the one must be false if the other is true. ‹x is a triangle›[8] is inconsistent with ‹x has four sides› because the one must be false if the other is true.)

The position that there is no philosophical truth is a common one. Every semester, at least ten students of mine openly advocate it. When such a student is asked why he holds that view, he always take one of three views:

(i) All legitimate knowledge is empirical knowledge; therefore, philosophy, being a non-empirical discipline, is no source of knowledge,

(ii) All legitimate knowledge is scientific knowledge therefore, philosophy, being a non-science, is no source of knowledge,

and

(iii) There is no truth where there is disagreement.

(i) is straightforwardly false, given that mathematics yields knowledge and that it isn't an empirical discipline. Also, as we've seen, any given empirical discipline makes assumptions that it cannot justify and that it, unless they're justified, that discipline fails to yield knowledge. So, supposing that, in (ii), the term "scientific knowledge" refers to the results generated by empirical sciences, it follows that (ii) is false. And supposing that, in (ii), that term refers, not just to empirical disciplines, but also to *non*-empirical disciplines (e.g. mathematics, philosophy), then what it says in no way implies that philosophy fails to yield knowledge.

(iii) is easily shown to be false. If one person disagrees with another, *at least* one of those people must be wrong. If Smith thinks that 2+2=5 and Jones disagrees, his position being that 2+2=6, then, even though each has the

wrong arithmetical belief, each is right to think that the other is wrong. If Smith thinks that 2+2=5 and Jones disagrees, his position being that 2+2=4, then, in addition to having the right arithmetical belief, Jones is right to think that Smith is wrong. So where there is disagreement, there *is* truth. Consequently, the alleged fact that there is more disagreement in philosophy than in other disciplines doesn't warrant the conclusion that there isn't such a thing as philosophical truth.

3.3. What philosophers agree about

There probably is more disagreement within the discipline of philosophy than there is within most other disciplines. But in philosophical disagreements have nothing to do with the logical structure of discipline of philosophy and instead have to do with egos of the people involved in this disagreements. The problem isn't that there aren't agreed-upon methods of resolving disputes. Nor is the problem that people disagree as to how those methods are to be applied. The problem is that, for whatever reason, philosophers practically never admit that they're wrong. They are *proven* wrong---and, when this happens, they usually know it. But, when proven wrong, they respond, not by admitting that they're wrong or even privately changing their views, but rather by coercing people over whom they power into towing their line. Since people don't depend for their well-being on how philosophical disputes are settled, market forces don't weed out bad philosophers the way they weed out bad doctors and engineers. That said, there is nothing inherent in the discipline of philosophy that prevents disputes from being resolved: one must distinguish the discipline itself from the personalities of its practitioners.

Also, despite everything just said, there *is* considerable agreement among analytic philosophers, and this agreement often concerns foundational issues. For example, analytic philosophers, setting aside completely incompetent ones, are unlikely to disagree as to whether a given ***proposition*** expresses a legitimate rule of inference, and, given a pair of propositions, they're unlikely to disagree as to whether the one ***entails*** the other. P entails Q if P cannot possibly be true if Q is false. (Thus, ‹*x* is triangle› entails ‹*x* has three sides. ›) A proposition is what is affirmed by an utterance of a declarative sentence.[9] Further, given two propositions, philosophers are likely to agree as to whether the one *confirms* the other. (P ***confirms*** Q if, other things being equal, Q is less likely to be true if P is false than if P is true. Given that Smith clearly isn't a butler or chauffer, Smith's living in a big house and driving a Rolls Royce confirms, but does not entail, that Smith is wealthy. Why doesn't it entail it? Because Smith might have stolen the car and he might not be able to afford the house.) So, granting that there is more disagreement in philosophy than there is botany or astronomy, there is also a great deal of agreement in philosophy.

3.4 What philosophers don't agree about

That said, after ego-driven fakery is set aside, there is at least as much disagreement in philosophy as there is agreement; and even though, as we've seen, philosophers agree about some foundational issues, they disagree about others. For example, even though, when asked whether one proposition entails another, different philosophers tend to give the same answer, there is no consensus among philosophers as to *what it is* for one proposition to entail another. Indeed, there is no agreed upon answer to the question "what *are* propositions?" Nor, consequently, is there any agreed upon answer to the question "to what extent, and in what way, are the structures of propositions reflected in those sentences that we use to express them?" This last question, though seemingly arcane and technical is profoundly important, as we'll see in Chapter 3, the reason being that, in the answer to it, lies the answer, or a large part of the answer, to the question “in what respects does a person’s knowing a language enhance his cognitive abilities?”

In the pages to come, we'll encounter dozens of other equally important philosophical questions that have yet to be resolved. We'll also find that, when supplemented with a modicum of technical rigor, a willingness to listen to one's intuitions, however heterodox they may be, often yields plausible answers to questions that, despite centuries of trying, philosophers have failed to answer with anything other than brazen falsehoods.

4.0 How Psychological and Sociological Factors Thwart Philosophical Progress

I suspect, on average, students of biology at first-tier institutions are way ahead of their counterparts at counterparts at bottom-tier universities in respect of their likelihood to discover biological truths. But philosophical discoveries are only marginally more likely to be made by students at top-tier institutions than at bottom-tier ones. I made this claim in the last section, but didn't defend it. Now I will defend it. My reason for discussing this topic is that, by doing so, we'll have a chance to discuss philosophical disputes that we won't have occasion to discuss in the pages following this preface.

There are two facts that, once clearly stated, are likely to be accepted, since they are consistent with generally accepted truths about human psychology, and that make it clear that this hypothesis of mine is at least approximately correct. The second fact is by far the more important one.

Fact #1: Whereas what passes for mathematical/musical ability in institutional contexts *really* is mathematical/musical ability, what passes for philosophical ability in such contexts is more likely than not to be a mere mimicry of it. Why is this? Two reasons. First, even though an interest in philosophy is extremely widespread, genuine ability at it is extremely uncommon. Second, whereas the difference between somebody who has musical talent and somebody who lacks it is brazenly obvious to almost everyone, the only exceptions being those who are tone-deaf or otherwise severely cognitively defective, the difference between somebody who has philosophical talent and somebody who lacks it is invisible to people who do not themselves have it. Given these two facts, it becomes inevitable that the people in charge of institutions whose very purpose is to cultivate philosophical insight haven't a shred of it themselves and are therefore can't judge who has philosophical talent and who doesn't. When judging whether a given person is philosophically gifted, such a person is forced to rely on some belief of his to the effect that people who have philosophical ability are likely to have other characteristics whose presence, unlike that of philosophical ability itself, are easy to detect. Those characteristics are typically ones that conduce to institution-internal success (e.g. orderliness, discipline in regards to following instructions, complacency in regards to forming views of one's own views) but are seldom ones that have anything to do with a person's degree of philosophical acumen. As a result, philosophy professors are likely

to be people who would have made good bureaucrats, but don't make good philosophers.

Fact #2: 99 times out of a 100, a brilliant student of philosophy at Cambridge or Princeton will use every ounce of intellect that he has to defend the views of his professors, since his professional future is in their hands; and he will make every effort to smear a position, however plausible, that isn't in lockstep with those of his professors. He'll find ways to defend behaviorism, functionalism, modal realism, or whatever happens at the time to be the line one must toe if one is to move forward in the philosophical world. 99% of the time, people gifted with philosophical intelligence use it to rationalize philosophical views that, although clearly broken, happen to be held by the people who give out philosophy-professorships. I will now give several examples of this.

4.1 Case in point: Boris Kment on counterfactual conditionals

Boris Kment, a contemporary of mine who received his degrees from Oxford and Princeton and must therefore surely be extremely intelligent, takes it for granted that, given some true claim about what *would* have been the case hold in virtue of the things that happen alternative universes that exist as concretely as our own. For example, suppose that Smith has exactly two cars; and, with that assumption in mind, consider the statement:

(1) If Smith had three cars, he would have an odd number cars.

According to Kment, the meaning of (1) is:

(2) There exists an alternative universe that is just likes ours except that, in it, Smith has three cars; and in that universe, Smith has an odd number of cars.[10]

I disagree with Kment. In my view, the meaning of (1) is:

> (3) The statement that Smith has an odd number of cars follows from the statement that Smith has three cars. In other words, one is guilty of holding a self-contradictory view if one holds that Smith holds three cars but not an odd number of cars.

The negation of (1) is incoherent. That's how we know that cannot possibly have three cars without having an odd number of cars. But if Kment is right (1) is an empirical proposition, the veridicality of which can be ascertained only by taking a tour of every alternate universe where Smith has three cars. This isn't how we know that (1) is true, and Kment's analysis is therefore false. [11]

Question: Why does somebody of Kment's level of intelligence not bother to consider alternatives to his view or even to consider questions that, although presupposing the correctness of his view, concern anything other than minor technicalities?

Answer: Kment’s adviser at Princeton was David Lewis (1940-2000). Like Kment, Lewis held that (2) is the correct analysis of (1) and that all statements of its kind are to be analyzed along the same lines.[12] In fact, this view, which is known as “modal realism,” is the cornerstone of Lewis’s work. Lewis was a philosophical genius and, in most philosophical contexts, was unusually flexible. But he wouldn’t budge when it came to modal realism. For him, that was a line in the sand: accept it or get lost. Kment accepted it. And, ten years after receiving his doctorate, he still does. In that time, everything he's published has been simply repeated or elaborated on points that he made in his dissertation. That said, Kment is tenured at one of the top ten departments in the country and I've been told more than once that he's one of the two or three finest philosophical minds of his generation.

4.2 Wittgenstein’s ruinous legacy

But, when one looks at the philosophical legacy of Ludwig Wittgenstein (1889-1951), one finds even starker examples of the principle that philosophers use their intelligence, not to find the truth, but to defend orthodoxies and to discredit any truth that should happen to conflict with those orthodoxies.

Wittgenstein was the most highly ranked professor at Trinity College Cambridge, then the most prestigious philosophy department in the world, and his views became orthodoxies. Unfortunately, those views were not only false but idiotic. His major work is titled *Philosophical Investigations.* The main thesis of that work is, I quote, that “thinking is not a mental process.” His position was that to think is to engage in overt behaviors that need not have mental accompaniments. And he held that, should such behaviors have mental accompaniments, they are in no way enriched by them---those mental accompaniments do not add, or help, to add a cognitive dimension to those behaviors. Whatever is going on in the relevant person's head is, to use Wittgenstein's term, a "free wheel."

Thus, Wittgenstein denied that what a person says is to any degree determined by what his beliefs are. For, according to Wittgenstein, what a person says *is* what a person thinks; so there can’t be thought *behind* the speech. There is just the speech. The speaker’s psychology is irrelevant. In fact, Wittgenstein went so far as to deny that there was *anything* to the mind other than what that person’s body did.

So according to Wittgenstein, what determines what a person is thinking is not what that person is thinking. (It's not hard not to state this contention of Wittgenstein's in a way that doesn't make it sound like a joke.) Whether one is thinking this as opposed to that is determined by how one's body is moving, what noises one is making, and where (in the cultural, not the geographical sense of "where") those noises and body-movements are occurring. Wittgenstein’s view was that you are the mask that you wear.[13]

Wittgenstein thus held that cannot experience pain without *acting* as though one is pain and, more generally, that what we refer to as “experiences” partly or entirely *consist of* the behaviors to which they tend to give rise. In his view, it isn't just what we think that is determined by how our bodies interact with those of others in our culture: what we *feel* is also so

determined.

In denying that thoughts had any existence apart from their overt manifestations, Wittgenstein was obliterating the distinction between voluntary and involuntary behavior. A voluntary movement is one that is caused by an intention on the part of the agent. But intentions are thoughts, or at least involve them; so, since Wittgenstein held that there is nothing to a person's thoughts other than his body movements, he was effectively denying that there is any difference between involuntary movements (such as those involved in an epileptic fit) and voluntary ones (such as those involved in my typing this sentence).

Wittgenstein's view was that we're zombies who don't think or feel anything unless we wiggle and jiggle in culturally accepted ways. This is not a *summary* of his position. It *is* his position. If it sounds as though I'm ridiculing it, that's only because I'm stating it clearly.

To anyone who is not a hollowed out pseudo-person, these claims are brazenly absurd. I suspect that, in his attempt to describe the psyche, Wittgenstein was describing himself. For hundreds of pages, Wittgenstein insists that

> (i) Thought has no existence independently of overt bodily behavior,
>
> (ii) Thought has no existence outside of a cultural context: to think is to act (or, in any case, for one's body to move) in ways that are in keeping with the mores (or, to use Wittgenstein's term, "forms of life") constitutive of one's culture

And

> (iii) Although some noises (e.g. an utterance of "it's raining") are meaningful, there are no such things as meanings: meaningfulness is a characteristic of certain behaviors and ink-marks and noises. But meanings *per se* are non-entities.

In Chapter 8, we'll discuss Wittgenstein's arguments for these contentions. But (i) and (ii) are self-evidently false. (iii) is not self-evidently false; but it is

easily refuted, as we'll see in Chapters 1, 2, 3, 4, and 8.

Stated as directly as possible, Wittgenstein's three main views are:

(a) We're empty shells. There's nothing on the inside.

(b) We are the masks we wear.

(c) Nothing means anything.

4.2.1 Wittgenstein’s nihilism about mind and meaning an unwitting admission about his mind and his work

Speaking from a psychiatric, as opposed to a philosophical standpoint, (a), (b), and (c) perfectly articulate the outlook of the *psychopath.* A psychopath is not an insane person. Nor is a psychopath necessarily an evil person. A psychopath is an empty fraudster whose personality-architecture is far too rudimentary for him to even grasp what a value is, let alone have one. Psychopaths are weak, underdeveloped con-artists; they aren't "cool" super-villains.

That Wittgenstein believed (a)-(c) to describe the human condition suggests that he himself was an empty shell who valued nothing and saw no meaning in anything. It suggests, in other words, that he was a psychopath. In Chapter 27, it will be argued that he was in fact a psychopath. [14]

Let us relate these points to our thesis that philosophers are more likely to rationalize existing orthodoxies, no matter how obviously false, than to look for philosophical truths. During Wittgenstein's 25 years at Cambridge and for several decades after his death, many of the brightest philosophers in Britain treated Wittgenstein’s views as simple truths. And most of those who didn’t agree with *all* of Wittgenstein's contentions didn‘t dare concede that, on the whole, his thinking was basically sound. As a result, some capable philosophers advocated some views that obviously no merit.

For example, Norman Malcolm (1911), a hardcore Wittgensteinian, denied that people dream.[15] His argument: Dreams, supposing them to exist, are mental processes; and we know from Wittgenstein that we don’t have mental processes.

4.2.2 Gilbert Ryle: The first of Wittgenstein's many causalities

Another hardcore Wittgensteinian, Gilbert Ryle (1900-1976), said that the difference between voluntary and involuntary actions isn't that the former result from certain kinds of mental states that the latter do not result from; the difference, he said, is that voluntary motions are culturally more acceptable, and tend to look more elegant, than involuntary ones.[16] According to Ryle, what makes a ballerina's movements voluntary is that they look good (or we like them or they're culturally approved of, or some such); and what makes a person's projectile vomiting involuntary is that it doesn't look so good. (I don't know what Ryle would say about the odd person who voluntarily projectile vomits.)

Ryle also held that, if I can see how your body is moving, I know what's going on in your mind at least as well as you do. So if somebody wants to know whether you are having a migraine, they should ask me, not you, unless you happen to be looking at yourself in a mirror.

Ryle was not stupid. He graduated with a double first class degree from Oxford. And even though most of what he said is Wittgensteinian, and therefore false, one can tell from the way he said it that his was a subtle and penetrating mind---or would have been, had he not matriculated in a culture of Wittgenstein-worship.

4.2.3 Peter Strawson: The greatest of Wittgenstein's many causalities

Ryle was not a great philosopher; he was merely a potentially great philosopher. But his colleague at Oxford, Peter Strawson (1919-2006), *was* a great philosopher. Unfortunately, like other British philosophers of his generation, Strawson was very reluctant to question The Master. As a result, he committed several hideous gaffes. One example is his "ascriptivist" analysis of moral responsibility. Another is his "expressivist" analysis of the concept of truth.

4.2.3.1 Strawson’s ascriptivism

Give or take some irrelevant nuances, Strawson held that, if it's correct, the statement"

(a) “X should be praised for doing Y”

holds in virtue of the fact that:

(a*) X deviates from the way in which people usually behave, and it does so in a way that pleases us.

Strawson believed the same thing *mutatis mutandis* to hold of any other ascription of moral responsibility. So held that, if they are true, the statements

(b) “X should be blamed for doing Y“

and

(c) “X is responsible for doing Y, but does not on that account deserve either praise or blame”

hold in virtue of, respectively, the fact that

(b*) Y is counterstatistical in a way that we don’t like

and the fact that

(c*) Y is not counterstatistical.

If Strawson is right, X doesn’t even have to have intentions for any of these statements to be true. Strawson's position thus entails that intentions either don’t exist or don’t do anything and therefore might as well not exist.

Strawson's analysis of ascriptions of moral responsibility is obviously

false. Inanimate objects don’t have intentions; so we can’t be held responsible for anything. And it's only a person's behavior is the result of an intention of his that we hold him responsible for it. We don’t hold somebody with Parkinson’s responsible for his tremors.

If Strawson were right, a corpse that fell on Smith could appropriately be blamed for Smith's subsequent death. But we don't blame the corpse; we blame the prankster who threw the corpse off a nearby rooftop.

Strawson is saying that whether a given person has a given intention depends on whether we ascribe an intention to him. So he's advocating the antithesis of the traditional view, which I regard as self-evidently correct, that whether that ascription is correct depends on whether that person has that intention. Strawson's view turns the living into the dead and the dead into the living. Sad about grandma's death? No problem! Just ascribe an intention to her---that'll bring her back. Don't like your boss? No need to worry! Just stop ascribing intentions to him---that'll zombify him.

Strawson was a brilliant man. So why did he advocate this grotesquely mistaken view? Because Wittgenstein advocated it. But whereas Wittgenstein said obscurely and he provided no argument for it, Strawson said it clearly and provided a semblance of an argument for it. So Strawson's article wasn't redundant.

4.2.3.2 Strawson’s expressivism

According to Strawson, the expression “true” means nothing. “It is true that snow is white” has the same meaning as “snow is white.” In prefacing the words “snow is white“ with the words “it is true that”, one is doing in a polite way what one would be doing in an impolite way by shouting “snow is white” while banging one’s fist on a table: one is showing one’s collocutor how convinced one is that snow is white. So the term “true” helps one *express* one’s feelings about a claim: they give one a civil way of express one’s certainty as to its truth. But that’s all that word does.

Strawson's expressivism is no less misguided than his acriptivism. First of all, the word “true” is not redundant. There is no sentence that means that same thing as “everything Plato said was true” that doesn’t contain the word “true” or some synonym thereof. Suppose that $S_1 \ldots S_n$ is a complete list of everything every said by Plato. The sentence “everything Plato said was true” would not have the same meaning as “S_1 is true and S_2 is true and…S_n is true.” Supposing that S_{37} is the proposition that there are infinitely many prime numbers, the statement “everything Plato said is true, but Plato never said that there are infinitely many prime numbers”, though false, would be coherent. (That's why one would have to know about Plato’s life to know whether that statement was true or not.) But statement “Plato said that there are infinitely many prime numbers and Plato never said that there are infinitely many prime numbers” would contradict itself. (That's why one *wouldn't* have to know anything about Plato’s life to know whether that statement was true or not.)[17]

"What about Strawson’s claim that “snow is white” and “it is true that snow is white” are synonymous? Surely that's got to be right."

It is not. The first attributes a color to a substance. The second attributes a truth-value to a proposition; so the second, unlike the first, is only indirectly about snow. Consider the sentence: “it is probable that snow is white.” That sentence *clearly* concerns a proposition. Snow is either white or it isn’t. When you say that snow is probably white, you’re not saying that it’s vaguely white or that its whiteness is only semi-existent. You are saying that, in saying that snow is white one, what one is saying is more likely than not to be correct.

The meaning had by the expression “that snow is white” in “it is probable that snow is white” coincides with counterpart in “it is true that snow is white.” So the latter, though ***equivalent*** with "snow is white", is about a proposition about snow, and therefore isn‘t in any direct sense about snow. “Snow is white“, on the other hand, clearly does directly concern snow. Two statements are equivalent if neither can possibly be true if the other is false.

Strawson had precisely the sort of intelligence one needs to see that, for these reasons, “true” *is* meaningful. So why did Strawson say otherwise? Because Wittgenstein said otherwise.

Why did Wittgenstein say otherwise? An integral part of Wittgenstein’s philosophical system is that meanings don’t exist. Wittgenstein granted that some noises were meaningful and that others weren’t; but he denied that there are such things as meanings.

Why did Wittgenstein hold *this*? It's a corollary of his rejection of the presumption that a least part of what distinguishes my saying “it’s raining,” which is a genuinely linguistic act, from a coffee machine's accidentally making those the same noises, which is not a linguistic act, is that, in the first case but not the second, those noises resulted from intentions embodying certain beliefs (e.g. my belief that it's raining). A belief is necessarily a belief *that* such and such is the case: the object of a belief is always a ***proposition***. A proposition is what one affirms when one makes a statement (e.g. when one sincerely says "it's raining"). Since Wittgenstein didn’t believe in beliefs, he didn’t believe in meanings; and since he didn’t believe in meanings, he didn’t think that “true”, which denotes a property of meanings, can possibly have a meaning.

Of course, it isn't just propositions that can be described as "true." Occurrences of sentences can also be so described. But, in saying “Jimmy’s words are the truth," one is saying in an abbreviated that the thing Jimmy affirmed is true. And propositions are, by definition, things that are affirmed (and denied and questioned, etc.).

4.3 Wittgenstein’s expressivism and the erroneousness thereof

Wittgenstein dedicated the second half of his professional career to establishing the non-existence of beliefs and everything else that exist only if meanings do as well. This led him to make a number of demonstrably false claims.

One of such claim is that, if a person says “I am in pain”, he is no more *stating* anything about his mind than he would be by writhing in agony. Saying "I am in pain," according to Wittgenstein, is merely a way of *showing* that one is in pain. So it's meaningless, except in the same non-linguistic sense as a baby's sobbing.

Here’s one way of showing that this position is false. Suppose that, in response to your saying that I am in pain but am trying to conceal it, I say:

> (A) If I am in pain, then certain neurons in my brain are firing that wouldn’t otherwise be firing; since, with the help of modern technology, I can demonstrate to you that those neurons are not firing, it is *not* the case that I am in pain.

If Wittgenstein were right to say that “I am in pain” was as devoid of meaning (in the linguistic sense) as the pain-driven howling of an injured animal, then (A) would have the same meaning, or lack thereof, as:

> (B) If *aaargh!*, then certain neurons in my brain are firing that wouldn’t otherwise be firing; since, with the help of modern technology, it is *not* the case that *ouch!*[18]

But B is meaningless and A is meaningful. So they don't have the same meaning, and Wittgenstein is wrong.

4.4 John Austin: Wittgenstein’s final victim

This blunder of Wittgenstein’s led John Austin (1911-1960), a faithful follower of his to make a similar, albeit subtler, mistake. Here is Austin's reasoning[19]:

> (AA)[20] Let *S* be a case of my saying to a friend: "I promise to give you a ride to the airport" S is not affirmation of some pre-existing fact. The promised referred to in *S* is *S* itself. But S doesn't describe itself. No utterance describes itself (with a few contrived exceptions, e.g. "this statement is false"). But, supposing for argument's sake that *S does* describe itself, *S*'s function isn't to describe itself, but rather to oblige the person who uttered it to perform a certain act.
>
> For exactly similar reasons, if a preacher says: "I now pronounce you man and wife," he's not describing some pre-existing fact. In uttering those words, the preacher was *performing* a social function. His wasn't a descriptive, but *performative* utterance, as was *S*.
>
> Performative utterances are meaningful. *S isn't like* "glarbojibjibjabber." Therefore, it isn't in virtue of affirming propositions that performative utterances are meaningful. To this extent, Wittgenstein was right to deny that meanings must exist in order for statements to be meaningful.

Austin's contention is false. If I say, "grass is green," what I am saying is either true or false. There is some truth or falsehood that I'm affirming---some *proposition*, in other words. And if I say "if grass is green, then grass isn't colorless," I am saying of that proposition that some other proposition follows from it. If, when occurring on its own, "grass is green" didn't affirm anything, then "if grass is green, then grass isn't colorless" would be as meaningless as:

(i) "If glarbojibjibjabber, then grass isn't colorless."

Also, if embedding "grass is green" in a larger sentence changed what it meant, then "if grass is green, grass isn't colorless" wouldn't say that grass's not being colorless was a consequence of its being green.

By parity of reasoning, if *x* is an utterance of the obviously meaningful sentence:

(ii) “If I promise to give you a ride to the airport next Tuesday, you’ll kill me if I break my word,”

and *y* is an utterance of:

(iii) “I promise to give you a ride to the airport next Tuesday,”

then *y* must affirm some truth or falsehood. If it didn't, *x* would be as meaningless as an utterance of (i). But clearly *does* affirm some truth or falsehood. And the same is therefore true of utterances of (i)

Thus, the spuriousness of Austin's position is a consequence of the fact, about to be demonstrated, that embedding a sentence *P* in ‹if *P*, then *Q*› does not typically change *P*'s meaning. The fact that

(I) there is an even prime larger than two

is false can be deduced from that:

(II) If there is an even prime larger than two, there is a prime number that is divisible by one, itself, and two, which is not possible

is true.

But (II) *wouldn't* necessarily be true if the occurrence in it of (I) had a meaning different from its usual one. For example, (II) would be true if that occurrence meant that 4 is an even prime larger than two. More importantly, if that occurrence had a meaning different from its usual one, (II) wouldn't

mean what it currently does, and it wouldn't necessarily say anything relevant to whether there was an even prime greater than two.

In general, if *P* meant one thing on its own and a different thing in ‹if *P*, then *Q*›, it would become impossible state ***indirect arguments.***[21] (An indirect argument is one of the form: *if P, then Q; not-Q; therefore not-P*.) But, to Austin's discredit, such arguments obviously can be stated.

In conclusion, so far as the philosophers just discussed are representative of philosophers in general, first-class philosophical thinkers are more likely to use their intelligence to defend orthodoxies, even wretchedly misguided ones, than to look for the truth. Ryle, Strawson, and Austin were at the world's first or second highest ranking philosophy department. And yet that they not only advocated views that any ten year old knows to be false. But they didn't just advocate them: they spent years of their lives urging acceptance of them. So these high-ranking members of high-ranking departments advocated views that were 100% false. So no one, even at a bottom-tier institution, could possibly say anything that's more false than what these high-rollers said. Hence my view students and professors of philosophy at first-tier institutions aren't as likely as one would think to produce better work than their counterparts at bottom-tier universities.

Part 1 The Analysis of Analysis

Chapter 1 Analytic Philosophy as Logical Analysis

1.0 Philosophy as the analysis of the categories in terms of which we understand the world

The world isn't homogeneous. It is articulated into *events* bearing various *causal* and, more generally, *spatiotemporal* relations to one another. Many of these events involve more or less persistent *things* (rocks, trees, etc.). Some of these things have *minds.* Most animate (mind-having) beings have *sense-perceptions;* i.e., they see things, hear things, etc. Most percipient (perception-having) creatures have *beliefs.* Some of these beliefs are *true;* others are *false.* Some of the true ones are cases of *knowledge.* Some percipient creatures communicate with one another through the use of *language.* Mastery of a language makes it easier for one to communicate one's beliefs to others, and it also enhances one's ability to *reason.* Rational (reason-capable) creatures tend to make *value-judgments.* They judge one another's actions, and sometimes their own, to be *good* or *bad.* Creatures that make such judgments tend to regulate their behavior towards one another by means of systems of *law*, the supposed purpose of which is to ensure that such behavior satisfies the requirements of *justice.*

Event, space, time, cause, persistence, thing, mind, perception, belief, knowledge, language, truth, value-judgment, law, justice. Whatever we know, we know it through these categories (and others like them—the list is not complete). Philosophy studies these categories. It delineates their structures. This is its sole function. Thus, *philosophy is the discipline that delineates the structures of the categories in terms of which we think about the world.* For reasons soon to be disclosed, this is the same as saying that philosophy is the science of science.

1.1 Knowledge vs. meta-knowledge

Although it is only *through* the aforementioned categories that we can think about the world, they are seldom the *objects* of thought. The result is that, although we're adept at using them, we know little about them. But for them, we are blind. And yet we are all blind to them. We know much, but have little idea what it is that we know. It is the philosopher's job to supply us with this missing knowledge.

For example, we are excellent at distinguishing linguistic from non-linguistic behavior, which suggests that, at some level, we know what conditions a creature's behavior must meet if it is to embody knowledge of a language. But when asked to make these conditions explicit, we find we can do so only with great difficulty and only with partial success. So even though we are good at knowing who knows a language and who does not, we don't know what it is that we know by virtue of knowing this.

The same thing ***mutatis mutandis***[22] is true of each one of these categories. We're excellent at distinguishing between moral and immoral behavior. We know that rape is immoral and that donating money to charity (for selfless reasons) is moral. But when asked to identify the principles embodied in these pedestrian and uncontroversial judgments, we have trouble producing theories that don't distort them.

In general, it's hard to identify the principles that guide our thoughts. Self-understanding isn't the mind's primary function. Nor could it be. The idea of a mind that thinks about nothing other than itself is an ***incoherent*** one. Thus, any case of self-awareness, and therefore of self-understanding, is necessarily derivative of, and for that reason of lesser quality than, some other more fundamental sort of understanding.

Being the discipline whose purpose is to delineate the structures of these categories, philosophy has the very non-trivial job of identifying, in as clear and explicit a manner as possible, the conditions that a given thing must satisfy if it is to fall into any given one of these categories. So far as philosophy succeeds in this endeavor, it clarifies what we are doing by virtue of doing anything cognitive, be it making an observation or engaging in an extended piece of abstract reasoning. Philosophy is the analysis of the preconditions of all knowledge.

Just as the scientist's job is to make us aware of the structure of the

physical universe, so the philosopher’s job is to make us aware of our knowledge of the knowledge of the universe. Philosophy is thus the “science of science," as John Stuart Mill put it.

1.2 The relationship of philosophy to other disciplines

The philosopher is interested in the *laws governing the laws.* He doesn't want to know what *in fact* holds. He wants to know what it would even make sense to claim to hold. He wants to know the laws that the laws can't break.

Philosophical knowledge is meta-knowledge—knowledge about knowledge. The non-philosopher wants to know specifics. What happened? When did it happen? What did it cause? How did it cause it? The philosopher is interested in these questions *only* to the extent that knowing the answers to them helps him understand the categories (*cause, place, time,* etc.) underlying such knowledge.

The stranded motorist wants his car to work. He doesn't care what will get it to do so. The engineer *is* interested in this. But the engineer is not entirely innocent of the motorist's epistemic parochialism.

The engineer has no interest in what the laws of physics are *except* in so far as he must know them to create the right mechanisms. But the physicist wants knowledge of these mechanisms *only* to the extent it will give him knowledge of the laws embodied therein.[23]

The philosopher is to the physicist what the physicist is to the engineer, and what the engineer is to the stranded motorist. It isn't for its own sake that the philosopher wants to know what causes what and what mechanisms were involved. He is interested in this knowledge *only* to the extent that knowing this helps him understand *what it is* for one thing to cause another— *only,* that is to say, in so far as it helps him know what it is that one knows in knowing that one thing made another thing happen.

1.3 How is analytic philosophy different from non-analytic philosophy?

John Stuart Mill (1806–1873), the great philosopher and economist, said that he was an expert in but one science, that being the science of science. What he meant was obviously similar to what we've been saying, and it was similar to what the great philosopher and mathematical logician Gottlob Frege (1848–1925) meant when he said that logic studies not the laws of nature but the "laws of the laws of nature." Whether Frege was right depends on what exactly one means by "logic." But if by "logic" he meant "philosophy," then Frege's dictum was spot-on.

Frege is often described as the first analytic philosopher. Michael Dummett (1925–), an exceptionally capable contemporary philosopher of language, said that "analytic philosophy is post-Fregean philosophy." What does Dummett mean?

With some exceptions, pre-Fregean philosophers thought they were studying the most general features of the actual world. They thought that, like the botanist, they were in the business of saying how the world is, the only difference between their work and the botanist's being that theirs is concerned with more general features of reality than the botanist's.

Frege showed that this is wrong. Any interest that botanists have in plants that *might* exist, but don't, is subordinate to their interest in what plants *actually* exist. Like all scientists, they are interested in what there *could* be only to the extent that it helps them figure out what there *is.* But with philosophers, it's the other way around. Any interest they have in the actual is subordinate to their interest in the possible. Unlike non-analytic philosophers, *analytic* philosophers *do* figure out what there could be by analyzing statements. Statements that make sense are those that can be true, and statements that don't are those that can't. Therefore, statements that make sense describe possible realities and those that don't, don't.

But what exactly did Frege do?

2.0 Frege's key insight: Logical form ≠ grammatical form

Frege's legacy to philosophy can be summed up thus:

> (FL) When people have an obviously correct belief that seems to have an absurd consequence, they should ask themselves whether that absurdity really *is* a consequence of that belief. But they frequently don't. Instead they accept the absurdity and, in order to make this mistake of theirs work, they develop *ad hoc* hypotheses as to the nature of reality that undermine the integrity of their own belief system.

2.0.1 FL explained

Consider the statement:

> (SC) nothing is a square circle.

SC is true. Everybody knows this. But what does SC say? Judging by its grammatical similarity to "Smith is a very capable lawyer," which attributes a certain property (that of being a capable lawyer) to a certain individual (Smith), SC would seem to say that a certain entity has the property of being a square circle. Presumably, the entity in question is some non-entity. If this presumption is correct, SC says that:

> (SC*) some non-entity, some featureless un-thing, is a square circle.

But SC* is doubly incoherent. If *anything* is a square circle, then SC is false—it being irrelevant how much of a cipher the entity in question is. Second, the very idea of a non-entity is an absurd one. (It makes no sense to say that there exists a non-entity, since anything that exists is, for that very reason, an entity.) But it is hard to find a layperson or scholar who, when asked what SC means, comes up with anything substantively different from SC*.

Given what a rank absurdity SC* is, we can't accept it, even though it seems to be an obvious consequence of SC. But we can't reject SC, since it's an obvious truth.

To get out of this jam, we need only reflect a bit on what our words mean. If you say that nobody likes Larry, you're not saying that some un-person likes Larry. You're saying that if you gathered together all the people who liked Larry and put them in an otherwise empty room, that room would remain empty. Which is the same as saying that if you put all the people in existence in an otherwise empty room, there would be no things that liked Larry in that room.

Thus, what "nobody likes Larry" means is not that some un-person likes Larry, but instead that the set of people who like Larry is empty or, alternatively, that the set of people doesn't overlap with the set of things that like Larry.

SC is to be understood along similar lines. If you put all the square circles in existence in an otherwise empty room, that room would remain empty. This is another way of saying that, if you put all the square things in an otherwise empty room, there would be no circles in that room. So SC says, not some un-thing is a circle, but that:

> (i) the set of things that are both circular and square is an empty one

or, alternatively, that:

> (ii) the set of circles doesn't overlap with the set of squares.

(i) is a way of saying that any given thing lacks the property of being both a circle and a square. Alternatively, it's a way of saying this property doesn't have any instances—that it's *uninstantiated.*

(ii) is a way of saying that anything having the property of being a square lacks the property of being a circle. Alternatively, it's a way of saying that these two properties don't have any instances in common—that they are not *coinstantiated.*

(i) and (ii) thus attribute properties to properties. They say, respectively, that the property of being both a square and a circle is uninstantiated and that the properties of being a square and of being a circle are not coinstantiated.

Since (i) and (ii) are just different ways of saying what SC says, the latter doesn't make the absurd statement that some non-entity is a circle, and it instead makes the innocuous statement that the set of squares doesn't overlap with the set of circles or, alternatively, that the property of being a square circle is uninstantiated.

What this shows is that, in at least some cases, philosophical insight is acquired not by doing para-science, but by analyzing meanings—not by positing entities, but by clarifying statements.[27]

Analytic philosophy is philosophy that is driven by acceptance of FL. Let us now expand on FL and make it clear why it is true.

2.1 The wrong way to react to grammatical surface-structure

In respect of its grammatical form:

(JS) “John smokes”

is comparable to:

(MS) “Mary smokes.”

JS attributes the property of being a smoker to John; MS attributes that property to Mary.

In respect of its grammatical form, each of JS and MS is comparable to:

(LJ) “Larry juggles”

and also to:

(JJ) “Jane jogs.”

Each of these sentences says of some individual that he or she has a certain property. The obvious inference to make is that *any* sentence grammatically comparable to any one of these four sentences says of some individual that he or she (or it) has some property.

In light of these points, consider the sentence:

(SS) Someone smokes.

Given what we just said, the obvious thing to say about it is that it attributes the property of smoking to some individual. But which individual could that be? Which individual does “someone” pick out?

“It picks out an ambiguous person,” said one logician.[28] But this answer is no good. Words are ambiguous, not people. “Bank” is ambiguous, since it has two meanings. But I’m not ambiguous, and neither are you. And if *per impossibile* there *did* exist some ambiguous person—some blank, featureless shell of a person who was picked out by “someone”—SS would unambiguously say of *that* person that he or she smoked. But there clearly isn’t any *one* person

to whom SS attributes the property of smoking. This is easily shown.[29] If Smith smokes, the sentence: "someone smokes but Smith does not" will be *false.* But it won't be ***self-contradictory;*** it won't be like "Smith smokes but Smith does not smoke." (For a statement to be self-contradictory is for it to bear two mutually opposed meanings.) Of course, there isn't anything special about the name "Smith," what we just said could have been said in connection with Jones or Brown or any other expression that refers to some individual. Thus, there is no individual *N* such that it is self-contradictory to say that someone smokes but *N* does not. Therefore "someone" doesn't refer to anyone.

"But you've misunderstood my thesis" it will be said. "The word "someone" doesn't unambiguously pick out an ambiguous person. It is *itself* ambiguous. It refers to Fred and Ethel and Mary. It refers to all people indifferently."

That's false. "Someone" isn't ambiguous; it isn't like the word "dumb." SS has one meaning, unlike "John is dumb," which could mean either "John is unintelligent" or "John is mute." Also, if "someone" were *ambiguous* between "John" and "Ethel" and so on, then, depending on the circumstances, it would be synonymous with "John smokes" or "Ethel smokes." And in that case, "someone smokes, but John does not" would sometimes have the same meaning as "John smokes, but John does not smoke," in which case it would be self-contradictory. But, as we just saw, "someone smokes, but John does not" is not self-contradictory under any circumstances. If John *does* smoke, it will be *false*, but it won't be self-contradictory.

Also, it isn't clear what it means to say that "someone" picks out everyone "indifferently." But if, as is surely the case, picking out everyone indifferently involves picking out everyone, then it's just wrong to say that "someone" picks out everyone "indifferently" (or in any other way). For if it did, "someone" would have the same meaning as "everyone," which it doesn't.

2.2 The right way to react to grammatical surface-structure

What all this shows is that "someone" isn't in the same category as "John" and "Ethel." It doesn't function in the same way, even though, given its *grammatical* function, one would expect it to. SS is obviously true. The wrong way to react to that fact is to twist reality to make it conform to our assumption that "someone," being grammatically comparable to "John," must refer to something. If we take that path, we must say that it refers to an "ambiguous" or "non-specific" person, given that it obviously doesn't refer to anyone specific. But then we're stuck with the absurd thesis that there is some non-specific entity in the world— that there exists something that isn't identical with any particular thing and therefore isn't identical with anything.

The right way to react to it is to think more deeply about what SS is really saying. Frege did this, and he solved the puzzle. SS's ***logical form*** diverges from its *grammatical* form. In terms of what its grammar suggests that it is saying, SS is indistinguishable from sentences that attribute the property of smoking to specific objects. In terms of what it really is saying, it is evidently very different from such sentences.

What SS is saying, as Frege made clear, is that the characteristic—or as analytic philosophers put it, the ***property***—of being a smoker is instantiated. SS is making a statement, not about some non-specific individual, but about a very specific property, and it's saying of this property that it's ***instantiated.*** (For a property to be instantiated is for there to be an ***instance*** of it. An instance of a property is something that has it. You are an instance of, and therefore instantiate, the property of being human since, being human, you have that property.)

There are properties that nothing has. Nobody has run a three-minute mile. Given any individual, *x*, it is false to attribute the property of being a person who has run a three-minute mile. Thus, there are no instances of that property. It is uninstantiated. The grammatical form of[30]:

(TM) "nobody has run a three-minute mile"

is just like that of:

(JTM) "John has run a three-minute mile."

JTM clearly attributes that property to an individual. That isn't what TM does. TM says of that property that it *can't* be attributed to anyone. So TM says that:

> (UP) [31] the property of being a person who has run a three-minute mile is uninstantiated.

Notice that UP's grammatical form is the opposite of TM's. The grammatical subject of TM is "nobody"; the grammatical predicate is "has run a three-minute mile." By contrast, the grammatical subject of UP is "the property of being a person who has run a three-minute mile," which corresponds to "has run a three-minute mile," and the grammatical predicate of UP is "is uninstantiated," which corresponds to "nobody."

If one were to take TM at face value, and were thus to assume that its meaning paralleled that of JTM, one would have to say that it attributed the property of being a person who has run a three-minute mile to some un-person. But if this is what it said, then in order to be true, somebody—namely, this unperson— *would* have had to have run a three-minute mile, in which case TM would be false. When we align TM's real meaning with its grammar, we don't have to swallow this rank absurdity. For UP says of some very much existent property (that of being a person who has run a three-minute mile) that *it* has a certain very much existent property (that of being uninstantiated). So we get the right result if we do linguistic analysis. We get the wrong result if we para-scientifically posit new *entities* in order to account for the datum that TM is true.

The same thing is true of:

> (NS) "Nothing smokes."

Pre-Fregeans said that NS says of some non-thing—some blank entity—that it smokes. So, supposing that in 500 years nobody smokes, and that NS is therefore true, it will be in virtue of the fact that some blank entity is around that is smoking.

But if there is such a thing, and it's smoking, then NS is *false.* As long as one thing smokes, NS is false. It doesn't matter how blank or otherwise deficient that thing is.

Frege straightened this all out. By obvious extensions of what we just said,

NS's real meaning is:

(NS_2) *The property of being a smoker* **is uninstantiated.**

NS_2's grammatical subject (the italicized expression) corresponds to NS's grammatical predicate; and NS_2's grammatical predicate (the bold-faced expression) corresponds to NS's grammatical subject. So when we align NS's logical form (what it actually says) with its grammatical form (what, given its grammar, it appears to say), we no longer have to say, completely absurdly, that NS attributes to some non-specific individual.

Similar remarks hold with respect to SC. That statement says of two properties that they aren't coinstantiated. It doesn't say of some non-entity that it is both a square and a circle.

Philosophical puzzles are solved by making it clear what statements mean. This is the basic tenet of analytic philosophy. Philosophy explains by clarifying statements. Science explains by positing entities. Science posits entities that are not themselves directly encountered but that, if assumed to exist, would account for phenomena that *are* directly encountered.

It used to be thought that philosophical explanation was to be understood in the same way—that philosophical progress was to be made by positing entities that are not themselves directly known but that, if assumed to exist, explain what *is* directly known. This is not the case. Philosophy isn't para-science. Philosophy is conceptual analysis. To make a philosophical discovery is not to discover a new entity; it is to make explicit a previously unrecognized implication of an existing belief. Philosophy is explication; it is the clarification of the statements that we accept but whose depths we haven't yet fully fathomed.[32]

3.0 Analysis vs. ontogenesis

In the works of analytic philosophers, one sometimes comes across the word “ontology,” which, etymologically, means “the study of being” (“ontos”= “being,” “logos”= “study”). These days the word “ontology” is usually used to refer to a given philosopher’s beliefs as to what exists. Some philosophers don’t believe in non-spatiotemporal entities. So such entities don’t belong to their “ontology.” Since I do believe in them, they do belong to my “ontology.”

Analytic philosophers are, almost by definition, ontologically very conservative. In other words, they don’t want to grant existence to anything whose existence hasn’t been demonstrated beyond a shadow of a doubt.

Pre-analytic, pre-Fregean philosophers were ontologically very liberal. Consider the statement:

(TP) No person is over 20-feet tall.

Pre-Fregean philosophers took TP to say that there existed some non-person who was over 20-feet tall; and, attempting to validate this analysis, they engaged in a great deal of spurious ***ontologizing***. (To ontologize is to posit an entity, or class of entities, not previously believed to exist.) Frege showed that this ontologizing, in addition to being futile, is unnecessary. TP says that a certain property (that of being a person who is over 20-feet tall) has another property (that of not being instantiated). No 20-foot tall (un)person need be postulated. No ontologizing need be done. By contrast, non-analytic (pre-Fregean) philosophers often countenanced bizarre and even logically impossible entities to cover up deficits in their analyses.

3.1 When is it appropriate to ontologize?

As a general rule, analytic philosophers do not try to solve philosophical problems by ontologizing. They try to solve them by clarifying statements. In *some* cases, the results of a successfully clarified statement *do* demand that we grant existence to something to which we'd otherwise deny existence. But, whenever this happens, the "entity" being posited is *never* a denizen of the spatiotemporal world; the thing that's posited is never a person or a table or a mountain or a monster. It's always an abstract object of some kind, and it's posited *only* because, were it not to exist, it would be impossible to account for the truth of obviously true statements.

For example, given the premise that Bob and Sally are both humans who are intelligent, it follows that there exist characteristics—or, to use the word preferred by analytic philosophers, ***properties***[34]—that Bob and Sally have in common, and from this it follows that properties exist.

Given that properties exist, are they identical with spatiotemporal entities (i.e., with things that have locations in space and time)? Obviously, *instances* of properties at least sometimes exist in space-time. Bob and Sally, both instances of many properties, exist in space and time. But although you may encounter many instances of intelligence, you'll never encounter intelligence *per se* and it would make no sense to assign any spatiotemporal location to it. Attempts to rebut this argument are doomed to fail, as we'll see in Chapter 2. Thus, unless there exists a property that Bob and Sally have in common, the inference from "Bob and Sally are both intelligent and kind" to "there is a property that they have in common" is illegitimate. Since that inference clearly is legitimate, such a property must exist.

Both every day and scientific thought are replete with inferences of this kind. Indeed, there would be no thought, let alone scientific thought, were it not for such inferences. Consider the inference from:

> (*) There is no actual or possible situation in which object *x*'s behavior differs from object *y*'s.

to:

> (**) *x* and *y* have similar microstructures; i.e., there is some

microstructural property that they have in common.

Were it denied that properties exist that inference would be invalid, and so would every other inference, the result being that “thought would grind to a halt,” to use Laurence Bonjour’s expression. So not only do property-*instances* exist: properties themselves do as well.

So far as they give any thought to the matter, most people hold that properties *don’t* exist, being nothing more than fictions created by philosophical chicanery. (When I applied to graduate school, my application-essay was an attempt to prove that, although property-instances obviously exist, properties *per se* do not.) But we’ve already seen one reason why such a view cannot possibly be correct, and we’re now about to see another.

There would be no truths unless there existed properties.[35] (Not property-*instances*, mind you, but properties *per se*.) Every truth is to the effect that some object has some property. There are apparent counter-examples to this. (For example, “nothing is a square circle” seems not to attribute any property to anything. But this is not the case, as we’ve already seen. In Section 4.9 of the present chapter, it will be shown that there cannot possibly be any actual counterexamples to this contention.) There would not be thoughts---or at least no correct ones----unless there existed truths. Thus, there would be no correct thoughts if there were no properties. Thus, lest we take the self-undermining position that all thoughts are false, we must posit the existence of properties whose existence we’d otherwise be inclined to deny. We must *ontologize*, in other words.

In Chapter 2, we’ll consider each of the many attempts to prove that properties do not exist; and we’ll also consider many attempts to prove that, although properties *do* exist, they have no existence apart from their own instances. We’ll see that all of these attempts fail. We’ll also see that, given a correct understanding of what properties are, any instinctual misgivings we have about positing their existence melt away.

3.1.1 When is it appropriate to ontologize? (Continued)

Even though both scientists and philosophers ontologize, they don't do so for the same reason. When scientists ontologize, it is in order to "model empirical data." It is in order to explain phenomena that, so they feel, they are not otherwise able to explain. Individuals $x_1 \ldots . x_n$ come down with various symptoms $s_1 \ldots s_m$. These symptoms must be explained. The causes of these symptoms must be identified. What we know about medicine and biology puts limits on what these causes might be, and so does what we know about the specific circumstances under which the relevant individuals fell ill. Given these two bodies of data, along with some principles of a generally logical nature, the most reasonable conjecture is that the source of the problem is a virus having such and such characteristics**.**

When analytic philosophers ontologize, they do it *only* when there is no other way of demonstrating the legitimacy of some indisputably correct form of inference, and they never posit anything non-spatiotemporal in the process. When scientists ontologize---that is, when they posit previously unknown entities---it is in order to identify the *causes* of states of affairs whose causes are still known. When scientists ontologize, it is in order to reconcile our knowledge of raw data (e.g., that Individuals $x_1 \ldots . x_n$ come down with various symptoms $s_1 \ldots s_m$) with our presumptions as to the causal structure of the universe---with, that is to say, our presumptions as to what states of affairs other states of affairs are likely to bring about. In Chapter 2, we'll see (i) that it is necessary to posit the existence of non-spatiotemporal entities and (ii) that, in positing such entities, one is no way deviating from either common sense or empirical science: one is simply make explicit what they both implicitly assume.

Whereas scientists ontologize in order to reconcile raw data with what we believe to be the principles governing redistributions of mass-energy, philosophers ontologize in order to show the legitimacy of the forms of inference underlying the specific inferences made by scientists.

It will be easier to see how the jobs of the scientist and the philosopher differ if we revisit the case of the medical researcher who identifies the Ebola virus as the cause of so and so's symptoms. That medical researcher is making a number of assumptions, all of them incapable of being questioned outside of a purely theoretical context by a sane human being, but not a single

one of them admitting of any obvious, strictly logical justification. Those assumptions include, but not are limited to, the following:

(i) The testimony of his senses can generally be trusted.

(ii) No knowledge concerning the external world or, therefore, concerning the causes of illness can be learned except through sense-perception.

(iii) Nothing happens for no reason. Everything has a cause. Nothing comes from nothing.

(iv) Even though things change, the structures within which they change tend to stay the same. People lose their health and then regain it; diseases come, go, and then sometimes come back in new, more virulent forms. But these changes happen within stable frameworks. Supposing that they change at all, the laws of biology change slowly, and they certainly don't change with every new case of illness or recovery. And even if, over a long period, the laws of biology do change, the laws of chemistry in terms of which those laws are to be understood do not change, and neither *a fortiori* do the laws of physics in terms of which the laws of chemistry are to be understood.

(v) Good explanations try to reconcile our presumptions as to how the world works---as to what the laws are that govern it---with the testimony of our senses. Sometimes those reconciliations are not possible: Either we must throw out the raw data or we must throw out one or more theories of ours as to how the world works. (We may throw out raw data if we discover that the microscopes we were using were defective. We may throw out a theory if, having exercised due diligence, we simply cannot find reason to question the raw data.)

Most reasonable people would accept most of these principles, if not all of them. But there can be no observation-based or, therefore, experimentation-based way of determining whether or not any given one of these given

principles is false.

Consider the principle that the testimony of the senses can generally be trusted.[36] Though integral to scientific and everyday practice, observations couldn't possibly verify it. After all, no observation is to be trusted *until* the truth of (i) has been established.

What about by asking others what their sense-perceptions were? That would be to no avail. First of all, it is only through your own sense-perceptions that you have any knowledge of what sense-perceptions others are having. So unless your own senses can be trusted, you can't trust other people's testimony as to what their senses are.

In fact, to continue further in this same vein, unless you assume the truth of (i), you cannot coherently assume that there exist other people. Nor, therefore, can you assume that, if there are other people, those people, when speaking to you, are being either honest or truthful.

3.1.2 A historical aside: Wittgenstein's blunder

Incidentally, the preceding remarks make it clear why one of the most famous philosophical arguments of all time is a failure. Ludwig Wittgenstein argued that one must know what others say and do in order to know what one is oneself thinking. He argued, for reasons that we will examine in Chapter 5, that one is to trust one's own views as to what one is thinking, feeling, etc., only if those views have been seconded by other members of the community of which one is a part.

The flaw in this argument is obvious. If I can't trust my senses, I have no way of knowing what you believe to be going on in my mind. If my senses can't be trusted, I don't even know that you exist, let alone that you have these or those views concerning my mental contents. So Wittgenstein's position is rank absurdity.

Let us now examine (iv). This contention is notoriously hard to justify. We know nothing about the future unless we know the laws that connect the present to the future. (If I know that the vase will shatter, it's because I know (i) that the vase is about to collide with a hard surface and (ii) that, the laws of nature being what they are, objects relevantly like that vase break when they collide with objects relevantly like that surface. If I know (i) but not (ii), I couldn't possibly know that the vase was about to break.) Thus, a precondition for knowledge of what will be the case at some future time is knowledge of the laws governing events between now and that time. Therefore, I have no way of knowing what those laws will be. Since I must know those laws to know the future, I cannot know what they will be. I cannot predict what they will be without knowing the future. Therefore, I can't know the laws linking the present to the future and I therefore can't know the future.

Of course, we're obviously disposed to believe that the natural laws that *will* govern the world coincide with those that *have* governed it. But the past is evidence of the future only to the extent that the mechanisms that have governed past events can be predicted to govern the future. Since this cannot be predicted, as we've just seen, nothing follows as to what the laws of nature will be from what they have been.

Are we to conclude that we can't know the future? No! We're to conclude that there's a problem with the argument just given. What is that problem?

See Chapter 7.

3.2 Brentano and Meinong: the non-analytic method epitomized

According to Franz Brentano (1838–1917), the essence of the mental is *intentionality*. In this context, the word "intentionality" refers not to the property of being done deliberately, but to the property of being *representational*. So, in Brentano's view, for something to be a mental entity is for it to be representational and for something to fail to be a mental entity is for it to be non-representational. To be sure, there are non-mental things (e.g., deposits of ink or paint) that are representational. Utterances and ink deposits are representational, even though they aren't mental. But this isn't a threat to Brentano's position. Utterances (etc.) are representational only in a derivative sense. It's because *we* endow it with meaning that an utterance of "snow is white" is meaningful; in a world devoid of sentient beings, it would just be another noise. So Brentano's thesis is that for something to be mental is for it to be non-derivatively representational.

There is an apparent problem with Brentano's thesis. To be representational is presumably to represent something. Hallucinations are mental entities. But what does a hallucination of a pink elephant represent? A pink elephant? No—pink elephants don't exist. More formally, there is nothing *x* such that *x* is a pink elephant. *A fortiori* there is nothing *x* such that *x* is a pink elephant that is represented by some hallucination. Still, there is clearly *a* sense in which hallucinations of pink elephants and other non-entities are representational. How is this to be explained?

The solution to this problem lies in the fact that *perception is description.* If you're looking at an actual elephant, which we'll call "Larry," the information encoded in your visual ***perception*** isn't the effect that:

> (LP) *Larry* is standing over there, next to that tree, looking ill (etc.).

When you look at an elephant, or a person, or a rock, you don't *just* see that object. Seeing Larry involves seeing *a* thing having various properties—having a certain color, shape, size, position (relative to you), etc. So seeing Larry involves seeing that various properties are instantiated—that there is an instance, in a certain place, of a certain morphology, color, etc.

Perceptions of things aren't like sentences about them. In the *sentence:*

(LS) "*Larry* is standing over there, next to that tree, looking ill (etc.)"

Larry is represented by a single, ***semantically simple*** symbol (namely, "Larry").[But in no *sense*-perception of Larry is he represented by some simple, homogenous, non-composite cipher. So far as he is perceived, Larry is represented *as* having these or those properties. This means that seeing him involves seeing that these or those properties are instantiated.

It's not as though *in addition* to seeing an instance of a certain morphology, color, etc., you *also,* separately from that, see Larry. No, for you to see Larry at a given time *is* for you to see that, at that time, certain properties are instantiated in a certain place. Thus, the content of your perception of Larry is given by some ***existence-claim*** similar to the following:

(LC) there exists, over in that place at the present time, instances of such and such properties.

An existence-claim is any claim to the effect that some property is instantiated. Thus, "there are prime numbers" is an existence-claim, and sufficient since it says that the property of being a prime number has at least one instance.

Given any existence claim, anything having the requisite properties is said to *satisfy* it. So the number seven satisfies the just-mentioned existence-claim. The number two *uniquely* satisfies "there is an even prime," since (a) it satisfies it and (b) nothing else does.

When you look at Larry, your eyes are giving you an existence-claim. Given that Larry, and Larry alone, satisfies that existence-claim, he is the *object* of your perception. So given that Larry uniquely instantiates the property of being a thing in such and such a place that has such and such a morphology (etc.), Larry is the object (or, more likely, one of the objects) of your current perception. So, yes, your perception does represent Larry. But it represents him by way of encoding an existence-claim that he satisfies.

Given this, suppose that, the next day you have a hallucination that is experientially just like the ***veridical*** perception we've been talking about. (A "veridical" perception is an accurate one. "Veridical" is to perceptions what

“true” is to sentences. For some reason, perceptions are described as “veridical,” not as “true.”) So even though neither Larry nor anything that looks like him is in front of you, your visual experience is telling you otherwise. That hallucination thus gives you a *false* message. The message encoded in it will be similar to LC. That message is to the effect that in a certain place there is a thing having such and such morphological, chromatic, and kinematic properties. On this occasion, the message—the existence-claim—in question is false, the reason being that nothing satisfies it. (Yesterday, the message represented by your visual experience was correct, the reason being that the existence-claim in question *was* satisfied.) The important point is that, although it was hallucinatory, your visual experience encoded an existence-claim and was therefore representational. It wasn’t representational in the sense that there was something that it picked out. It was representational in the sense that it gave you a message, albeit a false one, and thus represented the world as being a certain way.

Your visual perception of yesterday, unlike your visual perception of today, encoded a *true* existence-claim. But your visual experience of today no more represents some non-existent entity than your sense perception yesterday. Your visual experience today has for its content a *false* proposition to the effect that *there is* a thing having thus and such properties. But just as the sentence “there does not exist a ten-foot tall man” does not, in order to be true, require the existence of a ten-foot tall man, so your perception doesn’t require the existence of an elephant before you.

The same thing is true of thoughts about the non-existent. When you think about some non-existent number—for example, an even prime greater than two—there isn’t some mathematically impossible entity that you’re cognitively locking onto. What’s going on is that you’re thinking some false existence-claim along the lines of: there is some number *n* such that *n* is larger than two, even, and prime.

Mental entities have ***propositions*** for their contents. When correct, those propositions *describe* existing things. When false, they don’t. But nothing non-existent or quasi-existent can be the object of a thought or perception. When we *describe* a thought or perception as having a “non-existent object,” what we mean, so far as what we mean is coherent, is that it has for its content an existence-claim that nothing satisfies.

But Brentano dropped the ball. He realized that hallucinations are, in some significant sense, representational. Wishing to reconcile this with the fact that

there are no pink elephants (etc.), he said that a hallucination of a pink elephant has a non-existent pink elephant for its object.

But that's absurd, since it's the same as saying: "there exists some elephant *x* such that *x* doesn't exist and such that what you are hallucinating is *x*."

In a failed to attempt to deal with this, Brentano distinguished different kinds of non-existence, and he used different terms to mark them ("inexistent," "non-existent," "un-existent," etc.) His pupil, Alexius Meinong (1853–1920), added another bogus category to this list—the category of "subsistent" entities. An entity "subsists" if it doesn't quite fail to exist, but doesn't quite succeed in existing either.

This entire approach is misguided. Brentano and Meinong were ontologizing when they should have been analyzing. Properly analyzed, hallucinations no more require the existence of non-existent existents than veridical perceptions. The same is true of thoughts about Bart Simpson, the Fountain of Youth, etc. (This is further discussed in Chapters 6, 8, 9, and 25.)

4.0 Empirical puzzles vs. philosophical puzzles

Some puzzles result from ignorance of spatiotemporal facts. My valuables start disappearing. I'm puzzled. I learn the relevant facts: Larry has been sneaking into my house and stealing my valuables. I'm no longer puzzled.

The puzzles that science deals with typically involve a failure to know all the facts. Typically, the problem is not, at least not in its entirety, that anyone is making erroneous inferences. At least part of the problem is that not all the facts are in. It isn't yet known that disease *X* results from an over-production of antibody *Y*. The reason it isn't yet known is that given the available data there isn't yet good reason to believe it. But once the data is in, it will be believed, and a cure will be forthcoming.

Of course, scientific puzzles seldom result *entirely* from a failure to have the relevant data. In most cases, scientific breakthroughs involve somebody's figuring out a new and better way to model already available data.

The pre-history of relativity theory vividly illustrates this principle. If a train rushes past you at a rate of 100 mph, and I rush past you at a rate of 70 mph, the train is traveling at a rate of 30 mph with respect to me. But if a light beam rushes past you at a rate of 186,000 miles/second, and I rush past you at a rate of 180,000 miles/second, the light beam rushes past *me* at a rate of 186,000 miles/second. For this reason, *there is no optical test for determining one's own state of motion*.[42] In other words, no matter how quickly you travel, you will not be able to detect any change in your velocity *relative* to that of a light beam. People and instruments not traveling with you will be able to detect changes in *your* velocity relative to that of a light beam, and you will be able to detect changes in *their* velocities relative to that of a light beam. But nobody can detect any changes in his own velocity relative to that of a light beam.

This deeply puzzling fact was established in 1879. But nobody had any idea how to explain it until, in 1905, Einstein put forth the Special Theory of Relativity. Einstein didn't cite any data that hadn't been available to the physics community for decades. Einstein's great innovation was of a conceptual nature. The facts were in, but he was the first to make sense of them.

Be all of this as it may, Relativity Theory is an empirical theory. It's based on observational and experimental data, much of which would be impossible to acquire except through carefully executed experiments. This data wasn't in

until 1879. So even though nobody came up with Relativity Theory before 1879, that fact can't be chalked up to the fact that nobody drew the right inferences. It is due, at least in part, to the fact that the necessary data simply wasn't available. (What *can* be chalked up to a failure to draw the right inferences is that nobody came up with Relativity Theory during the period from 1879 to 1905.)

In general, scientific breakthroughs have two components: (i) a strictly fact-based component (new raw data is acquired), and (ii) a purely conceptual component (already known data is modeled in a new and better way).

4.1 Empirical puzzles vs. philosophical puzzles (continued)

Unlike scientific puzzles, philosophical puzzles are not solved by generating new raw data. Philosophical puzzles are *purely* conceptual in nature and have no strictly factual component. Philosophical puzzles result, not from a failure to know the facts, but from a failure to draw the right inferences. They result, not from ignorance, but from confusion. In figuring out that:

(NS_2) *the property of being a smoker* **is uninstantiated**

is what is meant by:

(NS) "nothing smokes,"

Frege solved a number of outstanding philosophical problems. But Frege didn't make any new empirical discoveries. In fact, his work didn't involve him having access to *any* empirical information that wasn't available to *anyone* who knows what (NS) (or its German equivalent) means.[44] Einstein's work, by contrast, had a heavy empirical component: He was modeling facts that a layperson would know nothing about, and the same thing *mutatis mutandis* is true of any scientific discovery.

4.2. Philosophical analysis ≠linguistic analysis

Impressed by Frege's philosophical successes, many came to the conclusion that *all* philosophical puzzles are of a purely linguistic nature. The most famous, and also the most vehement, advocate of this view was Ludwig Wittgenstein (1889–1951**).** This thesis was the cornerstone of his work. He urged acceptance of it in practically everything he wrote during his long career. He said it before others said it, and he held onto it long after most of its erstwhile supporters rejected it, for reasons to be described forthwith. "Philosophical confusion begins when language goes on holiday," he said. Elsewhere[45] he said that all philosophical confusion "lies in a failure to understand the workings of our own language."[46]

4.2.1 The Spuriousness of Wittgenstein's position

There are many philosophical puzzles that don't have anything to do with language, and the solutions to these puzzles cannot be modeled on Frege's solutions to the puzzles discussed a moment ago.

Here's an example from ***epistemology.***[Nothing can sense-perceive anything that doesn't affect it. You couldn't see the book in front of you unless light-rays bouncing off it struck your eyes. You couldn't hear the chirping of a bird unless that bird's chirping caused air-disturbances that led to ear-disturbances. All knowledge of what is in space-time is rooted in sense-perception. (If it's in space-time, it isn't known unless it's directly perceived or evidence of it is perceived.) Taking it for granted that nothing that isn't in space-time can possibly be known, many contemporary epistemologists; e.g., Jerry Fodor[48], hold that one can't know of *anything* without being affected by it. Thus, such philosophers hold that:

> (JK) John's being aware of the fact that 1 + 1 = 2 involves his being on the receiving end of some causal process initiated by that fact.

I personally regard JK as being absurd in the extreme.[49] The fact that 1 + 1 = 2 isn't comparable to the fact that there is a book in front of you. Were mass-energy distributed differently, there *wouldn't* be a book in front of you. But 1 + 1 would equal 2 no matter how mass-energy were distributed. Thus, "1 + 1 = 2" says nothing about how mass-energy is distributed, and it therefore says nothing about the spatiotemporal world. The fact it describes must therefore be non-spatiotemporal. Nothing outside of space-time can bear any causal relation to anything. Therefore, JK is wrong.[50]

In any case, the dispute between those who accept JK and those who reject it has nothing to do with language. It's agreed what JK means. What isn't agreed is whether the thing that it's agreed to mean is true.

But the controversies surrounding NS *do* concern language; they concern the semantics of the word "nothing." Those who see "nothing" as being a referring term, like "Socrates," see NS as saying that some featureless entity smokes. Those who deny that "nothing" and "Socrates" belong to the same semantic category don't see NS as saying this. Therein lies the controversy. Nothing comparable to this holds in connection with JK.

Analytic philosophers *do* agree that the right way to figure out whether or not JK is correct is by carefully analyzing its meaning. But they also hold that it isn't to be solved through *linguistic* analysis. They're right on both counts.

It must be emphasized that, according to analytic epistemologists, this puzzle is of a logical, not an empirical, nature. It is to be resolved through statement-analysis *alone*, not through statement-analysis combined with empirical research. They're clearly right about this. JK says that John's knowing that 1 + 1 = 2 is ***inconsistent*** with his not being on the receiving end of a causal process initiated by that fact. For P to be inconsistent with Q is for it to be impossible for both P and Q to be true. What doesn't exist cannot be observed. *A fortiori* what *cannot* exist cannot be observed. Thus, observation cannot tell you that anything is inconsistent with anything. So there is no way for it to tell you whether JK is correct.[51]

No philosophical assertions are empirical. Philosophy analyzes the categories in terms of which we think about the world. It does this by saying exactly what it is that is ruled out by a given thing's falling into a given category—by, for example, a given thing's being an instance of knowledge. Since observation can't tell one whether one statement is inconsistent with another, philosophical assertions are non-empirical.

W.V.O. Quine is the only philosopher I know to hold that epistemological disputes are *not* to be settled by analyzing concepts. In his paper "Epistemology Naturalized," Quine said that epistemological questions are not to be answered in this way, and are to be answered in the same way as scientific questions; e.g., "what is the Moon's chemical composition?"

There are two points to make here. First of all, even though Quine was an analytic philosopher by trade, it doesn't follow that all of his views were those of an analytic philosopher; and in this context, I would suggest, he was not being an analytic philosopher. He was being an opponent of analytic philosophy. Second, Quine's view is incoherent. Natural science is based on empirical observation. Empirical observation cannot tell you that it itself is a source of knowledge. Therefore, it cannot be known through natural science that observation is a source of knowledge, and it must therefore be known in some non-empirical, and therefore non-scientific, manner. Jaegwon Kim (1993) makes a similar point.

4.2.2 Not all philosophical analysis is identical with linguistic analysis (continued)

Thus, not all philosophical puzzles are linguistic puzzles; and Wittgenstein was wrong to say otherwise. But didn't we ourselves say that philosophical analysis is the analysis of statements? Yes we did, and we were right to do so. To see why Wittgenstein is nonetheless wrong, we must distinguish sentences from ***propositions.*** Propositions are the things meant by sentences. "Snow is white," "schnee is weiss," and "la neige est blanche" all mean the same thing. There is some one proposition that is the meaning of each of them.[52]

A sentence is true or false depending on whether it has a true or false proposition for its meaning. Thus, when a sentence is described as "true," the property being attributed to it isn't the same as the property that is attributed to a proposition that is so described. For a sentence to be "true" is for it to encode a true proposition. But this isn't what it is for a proposition to be true, since propositions don't encode anything. In Chapter 3, we'll say what exactly it is for a proposition to be true. But the obvious answer, though imprecise, is the right one; namely, for a proposition to be true is for it to fit the facts.

The term "statement" is ambiguous; it has three distinct meanings. Sometimes it refers to propositions, sometimes it refers to the sentences used to affirm them, and sometimes it refers to the *act* of using a sentence to affirm a proposition. Wittgenstein didn't countenance the existence of propositions, and this obviously had a hand in his erroneously believing that philosophy is the analysis of *sentences,* when the truth is that philosophy is the analysis of *propositions.* Even Frege's analyses of sentences such as "someone snores" and "nothing is a square circle" fail to conform to Wittgenstein's conception of what philosophy is supposed to do, since, as we'll see in a moment, they're analyses of propositions that *involve* analyses of sentences and therefore are not themselves analyses of sentences.

One must know at least some English to understand the *sentence:*

(1) "John knows that 1 + 1 = 2."

But the proposition meant by that sentence can be grasped without speaking English; and one needn't know English, or any other given

language, to be able to analyze that proposition correctly.

But analyzing (1)—the sentence, not the corresponding proposition—*does* involve such knowledge. A sentence is an expression; it consists of nouns, verbs, etc. Analyzing (1) involves knowing the various grammatical nuances involved in its structure. Analyzing the corresponding proposition has nothing to do with anything relating to grammar or any other aspect of language. No such knowledge is needed to analyze the corresponding proposition.

Even though Frege's work inspired many to identify philosophy with sentence-analysis, Frege himself always made it very clear that propositions are not sentences and that although sentences are human creations, their meanings are not.[53]

The sentence "the moon is less massive than Earth" is a human artifact; it didn't exist until a few centuries ago. But the truth it expresses is in a different category. That truth exists independently of us. After all, the moon was less massive than the Earth before we came along, and it'll be that way after we sign out. For the same reason *mutatis mutandis,* the truth expressed by "the Earth is more massive than the moon" exists independently of our thoughts and deeds, even though that sentence itself is a human artifact.

4.2.3 Not all philosophical analysis is identical with linguistic analysis (continued): the nature of sentence-meanings (as opposed to sentences)

What are propositions? They're properties. For a proposition to be true is for the world to be a certain way. The proposition that Smith is in Richmond is true if the world is a certain way and it's false if it isn't. (If mass-energy is distributed one way, Smith is in Richmond; if it's distributed some other way, he isn't.) For a thing to be a certain way is for it to have a certain property. To be round is to be one way; to be square is to be some other way. To be round is to have one property; to be square is to have some other property.[54] Thus, the world's being a certain way is identical with its having a certain property. Since, therefore, the world's being a certain way is identical with some proposition's being true, propositions must be identified with properties and a proposition's being true must be identified with its being instantiated. Propositions are properties and truth is instantiatedness.[55]

It's widely thought that propositions are human creations. This is false. The world was a certain way before we were around; it will be a certain way after we're gone; and the way it is while we're around is up to us only to a limited extent.[56] It follows that, independently of our having any beliefs or, indeed, our doing or thinking anything, certain propositions are true; and it follows from this that propositions are not human creations. Sentences, on the other hand, are human creations. They wouldn't be around if it weren't for us. So sentential analysis is a very different thing from propositional analysis.

4.2.4 Not all philosophical analysis is identical with linguistic analysis (continued): Frege's accomplishments reassessed

But didn't we say that Frege's great accomplishment lay in his insights concerning *sentences*—in his seeing that a *sentence's* surface structure sometimes pulled part from its deep structure? And didn't we say that, for this very reason, Frege was the first *analytic* philosopher—that, as Michael Dummett put it, analytic philosophy is post-Fregean philosophy, the reason being that analytic philosophy is *statement*-analysis? Yes, we did say all this. And yet we just spent a lot of time saying how analytic philosophy is "statement-analysis" *only* in the sense of being statement-*meaning*-analysis; that is, proposition-analysis, as opposed to *sentence*-analysis. How are we to reconcile those various statements with one another?

The term "analytic philosophy" can be construed narrowly or broadly. Construed narrowly, analytic philosophy is post-Fregean philosophy in the sense that it directly flows out of Frege's work. Frege was interested in ***reference, quantification,*** the nature of ***logical truth,*** the extent to which it's possible to ***formalize*** intuitively valid inferences, etc. (These terms will be defined soon enough, if they haven't been already.) The term "analytic philosophy" sometimes refers to what is done by those people who write *about those very questions* and who, in so doing, *are taking what Frege had to say about them into account.* In other words, "analytic philosophy" sometimes refers to the ***philosophy of language,*** along with a related branch of philosophy, known as ***philosophical logic.***[57]

But Frege's work had a profound influence on philosophers who were working in areas that have no direct connection to language or logic. Frege showed that, by thinking clearly, systematically, and self-critically, one could make real headway on philosophical problems in which others had yet to make so much as a dent, despite hundreds of years of trying. "The devil is in the details," as they say. Pre-Fregean philosophers tended to disregard the details. Frege did not. He was a stickler for them. And it was partly, though obviously not entirely, for this reason that he was able to solve problems that his predecessors could not. The word "analytic philosopher" sometimes refers to the sort of philosophy done by people who aspire to approach philosophical problems in the same careful and clear-headed way as Frege.

Thus, there are analytic philosophers of law, analytic ethicists, analytic philosophers of religion, etc. What makes them analytic philosophers isn't

that they're talking about language or logic or any of the things that Frege talked about. It's that they believe the problems they're concerned with can be solved in the same coolheaded, logical way that Frege solved problems relating to language and logic.

But even when taken in the narrow sense, "analytic philosophy" (i.e., philosophical logic/the philosophy of language) is only misleadingly described as the analysis of sentences. Philosophers of language are interested in the sentence:

(NS) "Nothing is a square circle,"

only to the extent that by understanding it they will deepen their insight into concepts of a general kind (e.g., ***meaning, analytic truth, modality***). Linguists, on the other hand, are interested in those concepts only to the extent that an understanding of them will help them understand specific sentences, such as NS. Frege did indeed painstakingly analyze specific sentences. But he did so only because he knew that, by so doing, he could identify general logical principles. Thus, for Frege, his insight that logical and grammatical form pull apart was ultimately just a means to an end, the end being the identification of the actual nature of the bearing-relations that propositions have with respect to one another. So, yes, analytic philosophy is statement-analysis; and, yes, it was Frege's brilliant analyses of *sentences* that availed philosophers of the principles needed to analyze statements properly. But analytic philosophy, even the narrow sense of the term, is not itself sentential analysis.

4.2.5 Not all philosophical analysis is identical with linguistic analysis (continued): Wittgenstein—an introduction

Wittgenstein's two best-known works are the *Tractatus Logico-philosophicus* (TLP), which he completed in 1921, and the *Philosophical Investigations*, which he completed in 1949. In many ways, these works are antithetical to each other. But in both of them, Wittgenstein insists that philosophical problems arise when, and only when, sentences are misused and are solved when, and only when, it is made clear how they are being misused.

In the TLP, Wittgenstein contends that all philosophical problems result from a failure to understand the syntactic rules of the languages we use. Sentences that would, if meaningful, express philosophical propositions are in all cases ungrammatical nonsense and thus fail to say anything. All such sentences violate the ***syntactic rules*** of the languages to which they belong. (For the time being, "syntactic" may be taken to be synonymous with "grammatical." See Chapter 4, Section 3.3 for a definition of "syntax.") Because they violate these rules in subtle, easily overlooked ways, they aren't always seen for the abject nonsense that they are. But we mustn't let appearances deceive us. All such sentences are syntactically ill-formed and therefore devoid of meaning, and there would be no philosophical problems if people fully understood the syntactic rules of the languages they spoke. So far as philosophy has any legitimate function, it is to identify these rules, thereby heading off the syntactic blunders that lead to philosophical puzzlement.

In the *Philosophical Investigations* Wittgenstein says, just as he does in the TLP, that philosophical problems arise when language is misused. But in the *Investigations* he denies that such misuses involve violations of hidden syntactic rules, and instead says that such misuses consist exclusively in one's using words in ways in which they are not *ordinarily* used. Thus, Wittgenstein's position in the *Investigations* is that all philosophical problems can be quickly and definitively solved by looking at how words are *actually* used and, on the basis of the knowledge thereby obtained, ceasing to use words in deviant ways. What we think of as philosophical puzzles concerning knowledge, logic, and morality are puzzles about the words "knowledge," "logic," and "morality." Those puzzles are created by our using those words in non-standard ways, and they're solved by our ceasing to do so.

4.2.6 Is meaning identical with use?

In a moment, we'll evaluate Wittgenstein's (1922) contention that all philosophical statements are ungrammatical nonsense. Right now, let us consider Wittgenstein's (1958) contention that philosophical puzzles are dissolved by looking at how words are used.

This contention is incoherent on many levels. Given only the acoustical and morphological properties of its spoken and written occurrences, the word "knowledge" could mean anything.[58] So, supposing that it's the concept of knowledge that we wish to learn about, we can't possibly know that the word "knowledge" is the right word to study *unless* we know that it expresses the right concept. But we can't possibly know *that* unless we have some way of grasping that concept that doesn't involve our operating with that word. So we can't even act on Wittgenstein's exhortation that we study knowledge by studying how people use the word "knowledge" *unless* we have some way of grasping the concept of knowledge that *doesn't* involve that word. But if we don't need that word to grasp that concept, we don't need it to study it; and if we don't need it to study it, then Wittgenstein is simply wrong to say that one must study how it is used to understand the concept it expresses. Given any expression *E* and any concept *C*, an obvious extension of this argument shows that one can't learn about *C* by studying *E* unless one can grasp *C*, and can therefore examine it, *without E*'s help. Thus, Wittgenstein is simply wrong to contend that one learns about concepts by learning about how the corresponding words are used.

"But you've overstated what Wittgenstein is claiming," it will be said. "Contrary to what you allege, he wasn't saying that concepts are to be learned about *solely* by looking at how the corresponding words are typically used. He was saying only that knowledge of expression-usage would be a useful adjunct to some other, more important way of knowing about concepts."

If that's what Wittgenstein is saying, then he's conceding everything said in the paragraph preceding the last one.[59] To grant that there is any expression-independent way of grasping concepts is to grant that expression-usage is to be evaluated *in light of* the very conceptual knowledge that, according to Wittgenstein's thesis, one is supposed to acquire *through* the study of expression-usage.

But there's a problem with our argument. It assumes that, given a

meaningful expression (e.g., “knowledge”), there is some object that is its meaning. Wittgenstein rejects this assumption. He holds that there is no *entity* that is the meaning of “knowledge.” Wittgenstein’s position is, I quote, that “meaning is use.”[60] In other words, for an expression to have a given meaning is for it to be used in a certain way. An expression’s having a given meaning does *not,* in Wittgenstein’s view, involve there being some entity that is its meaning.

Before we evaluate this contention, we must make a few facts about it clear. Wittgenstein isn’t making the uncontroversial point that how expressions are used is a function, in part, of what they mean. Nor is he making the equally innocuous point that how an expression is used may have *effects* on what it means. (“Probable” used to mean “capable of being definitively established.” Enough people used it to mean “likely, but not certain”; and, for that reason, that’s what it now means.) He is saying that *what it is* for an expression to have a given meaning is for it to be used in a certain way. Here, in my words, not Wittgenstein's, is his argument:

> (WA)[61] According to some philosophers, there is some entity that is the meaning of “snow is white” and some other entity that is the meaning of “grass is green.” These entities are known as *propositions.* Propositions (if existent) are non-spatiotemporal entities.
>
> The very idea of such an entity is of doubtful coherence. And even if such entities do exist, they’re nothing to us. We can’t see them or touch them or otherwise have anything to do with them. So, even if they do exist, they have no role in human affairs. At the same time, “snow is white” and “grass is green” differ in meaning.
>
> So we must find a way of saying what it is for two expressions to ‘differ in meaning’ that doesn’t involve our positing meanings.
>
> This can be done. “Snow is white” isn’t used in the same way as “grass is green.” There are situations that prompt utterances of the one that don’t prompt utterances of the other.
>
> By the same token, if they *were* used in the same way—i.e., if there were no situation that prompted utterances of the one that didn’t prompt utterances of the other—then they wouldn’t differ

> in meaning. If the sensory stimulations that induced people to say "grass is green" coincided with those that prompted people to say "snow is white," there would be no significant sense in which they "differed in meaning."[62] Thus, two sentences S_1 and S_2 coincide in meaning exactly if they are used in the same way. So instead of identifying the meaning of "grass is green" with some non-spatiotemporal entity—with some entity whose existence is in doubt and whose role in human affairs, supposing it to exist, is also in doubt—we can just say that two expressions "have the same meaning" if they're used in the same way. An obvious corollary is that for an expression to have this as opposed to that "meaning" is for it to be used in this as opposed to that way.

Were WA cogent, Wittgenstein would be doing to meanings what Frege did to square circles. Frege got rid of square circles by showing that sentences appearing to require their existence (e.g., "square circles simultaneously have, and lack, uniform curvatures") are equivalent to sentences that don't (e.g., "the statement *x is a circle* entails *x has a uniform curvature* and the statement *x is a square* entails *x does not have a uniform curvature*"). Wittgenstein wishes to get rid of meanings by showing that statements appearing to require the existence of expression-meanings are equivalent to statements (about expression-usage) that don't.

This wish of Wittgenstein's embodies some very wrong views as to what linguistic expressions are. A burst of noise that doesn't have a meaning is just a burst of noise. A burst of noise is an *expression* only if it has a meaning. But if it already has a meaning, then how it is being used isn't what gives it its meaning.

Our knowledge of what words mean is what *guides* our linguistic behavior. Expressions have meanings. We know this. And that's why we use expressions in the way we do. I know what "hug" means; I also know what "discuss" means. That's why, when talking to my students, I say "I want to discuss the exam," and not "I want to hug the exam." If meaning were use, meaning couldn't *guide* use. But it obviously does.

To the extent that meaning *doesn't* guide use, people are *mis*speaking. Wanting to tell you that you're an absolute genius, I say "you're an absolute moron," since, despite my generally good command of the English language,

I wrongly think that “moron” means *genius*. To the extent that my uttering those words embodied a *failure* to know the actual meaning of what I was saying—to the extent that meaning *failed* to guide use, in other words—I *mis*spoke. And to the extent that my uttering those words embodied a knowledge of what they meant—to the extent that meaning *did* guide use, in other words—I spoke properly. This shows that, to the extent that utterances aren’t simply defective, meaning guides use and consequently that use is *not* constitutive of meaning.

Bearing these points in mind, let us revisit Wittgenstein’s (1958) contention that it’s a philosopher’s job, not to analyze concepts or meanings or other such alleged phantasms, but merely to take note of *when* words like “justice,” “knowledge,” and “truth” are used. What would it be to do this? The situations in which the word “justice” is used don’t necessarily have any observable characteristics in common with one another. It’s not as though people utter the word “justice” when, and only when, they’ve just been bitten by a cat. This isn’t to say that uses of the word “justice” cannot be correlated in *any* way with facts about the situations in which those uses occur. Such correlations clearly can be made. People are likely to use that word when they’re in classes concerning *ethics* or the philosophy of *law;* they’re likely to use it when they *believe* that they’ve been *wronged*; they’re likely to be used by politicians who want people to believe them to be *worthy* of holding office.

The italicized terms express extremely abstract concepts; and utterances of those words are no more capable than utterances of the word “justice” being correlated with observable facts about the contexts in which they occur. So, while a person’s decision to use the word “justice” (or “knowledge” or “law,” etc.) may obviously have a situational basis, there is no way to identify that basis *except* in terms of the very concepts that, according to Wittgenstein, are to be studied by figuring *when* those terms are used. In other words, any correct generalization as to when words like “justice,” “logic,” “knowledge” and “morality” are used will *itself* employ the very concepts that, according to Wittgenstein, knowledge of such generalizations is supposed to yield.

These reflections bring us face to face with an incoherence inherent in all of the different variants of the contention that philosophy is sentence-analysis. To speak isn’t just to make noises; it’s to make noises for the reason that one believes those noises to have certain meanings. I’m with my friend Larry. All of a sudden he starts convulsing and foaming at the mouth. I call 9-1-1 and say:

(i) “My friend is foaming at the mouth and convulsing; please send help.”

I did *not* say:

(ii) “I like pizza,”

or:

(iii) “Giraffes are friendly creatures.”

Why did I utter (i)? Why didn’t I utter (ii) or (iii)? Because I know that (i) has the right meaning and that (ii) and (iii) don’t. In general, speaking consists of making noises for the reason that one believes that, given existing semantic rules, those noises have certain meanings. If somebody makes a noise that he does *not* believe to be assigned a meaning by any semantic rule, he isn’t speaking. If, not believing that the noises I’m about to make are assigned any meaning by any semantic rules, I say “blurga dunga blurbo,” I am not speaking. I’m just making noise. Maybe the semantic rules of some language *do* assign them a meaning. That doesn’t matter. My making that noise wasn’t guided by my knowledge of such rules. So I wasn’t saying anything.[63] And as we saw earlier, I am *mis*speaking if I misidentify the meaning assigned by existing semantic rules to the noises I am making. So there is no *bona fide* speech where there isn’t awareness of semantic rules.

Thus, the analysis of sentences isn’t the analysis of noises. (In this context, take references to “noises” to be short for references to anything that can constitute the occurrence of an expression—e.g., hand-movements, patterns of light on a monitor, etc.) In and of themselves, noises aren’t speech, as we just saw. It is only when a noise embodies an attempt to follow a semantic rule that it constitutes speech. So the analysis of speech—that is, of spoken sentences—concerns noises *only* to the extent that they embody attempts to follow such rules; and for the same reason *mutatis mutandis*, the analysis of sentences *per se*, as opposed to their spoken (or written) occurrences, necessarily involves, if it doesn’t coincide with, the analysis of those rules.

Those rules are not themselves sentences. The semantic rule that assigns meaning to “snow is white” is not itself a sentence.[64] Given these points,

what might it mean to say that philosophy is sentential analysis? It *could* mean that philosophy is the empirical study of different languages; i.e., that philosophy is linguistics. But that clearly isn't what philosophy is. So it must mean that philosophy is the study of semantic rules *qua* semantic rules—that, in other words, it is the study of the *concept* of a semantic rule.

In that case, philosophy is also the study of those concepts that must be understood to understand the concept of a semantic rule. And there are many such concepts. Among them are narrowly semantic concepts such as ***compositionality, reference, quantification, force,*** and ***negation.***[65]

But many of these concepts aren't *only* of relevance to semantics. As we've seen, no noise constitutes a sentence-utterance unless it embodies an ***intention*** of a certain kind; and there is no linguistic behavior of any kind where people aren't intentionally following what they ***believe*** to be existing semantic ***rules.*** Some believe that semantic rules are ***conventions.*** Others believe that they are ***functions*** (in the mathematical sense)— assignments of meanings or ***truth-conditions*** to noises or to properties that are instantiated by noises (or ink marks, etc.). There is no way to figure out what semantic rules are without examining these concepts.

And even if these questions are side-stepped, and philosophers focus *only* on narrowly semantic concepts (e.g., reference, compositionality, etc.), philosophy ends up being the analysis, not of sentences, but of *concepts.* Thus, the thesis that philosophy is the study of sentences either collapses into the obviously false thesis that philosophy is the empirical study of language *or* into the thesis that philosophy is conceptual analysis (i.e., the analysis of the concepts in terms of which the world is understood). Thus, Wittgenstein's contention that philosophy is the study of sentences collapses into the very view to which it is meant to be an alternative (viz., that philosophy is the analysis of concepts, as opposed to expressions), and is therefore false.

Let us now turn to Wittgenstein's (1922) Tractarian contention that philosophical statements are ungrammatical nonsense.

4.2.7 The Tractarian contention that philosophical statements are syntactically ill-formed nonsense

The sentence:

> (a) "One can be aware of the fact that 1 + 1 = 2 without being causally affected by that fact,"

appears meaningful. So does the sentence:

> (b) "There could in principle be a legal system that failed to embody any morality at all."

But in the TLP Wittgenstein says that this is an illusion. Here is the viewpoint underlying this bold contention:

> (TA)[67] All philosophical statements are ungrammatical nonsense, and all philosophical problems would vanish if we spoke grammatically. "All philosophical problems belong to the same class as the question whether the good is more or less identical than the beautiful."[68] Thus, all philosophical statements belong to the same class as:
>
> (BG) "the good is more or less identical than the beautiful."
>
> BG is obviously meaningless. The reason for this is that it's syntactically ill-formed. If we produced only syntactically well-formed sentences, we wouldn't produce nonsense like BG. Since all philosophical statements are in the same class as BG, we'd never produce any sentences whose merits it was the job of philosophy to determine.
>
> Were the logical forms of (a) and (b) brought into alignment with their grammatical forms—in other words, were their actual meanings reflected in their grammar—they'd be ungrammatical since they have *no* meanings.

Although TA is a paraphrase, not a quotation, the part in quotes *is* an exact quotation from the TLP.[69]

What Wittgenstein is saying isn't confined to the defensible claim that philosophical problems are to be solved by clarifying statements. Wittgenstein is making two additional claims. First, no sentence that appears to make a philosophical statement means *anything*. Second, it's *only* because they're syntactically ill-formed that such sentences are meaningless—there is no other reason.

Let us evaluate these claims. Though ungrammatical, the sentence:

(MH) "Me and Herby play tennis every day, and me always win because Herby not in good shape,"

is perfectly meaningful. Ungrammatical statements are often meaningful. Thus, BG's failure to mean anything cannot be entirely blamed on its being ungrammatical.

What's the real problem with BG? The expression "more identical than" is obviously supposed to function in the same way as relational expressions, like "more important than" or "identical with." But there is no relation that it picks out. Though it *consists* of English expressions, the expression "more identical than" is *itself* no more a part of the English lexicon than "blurga derba gurb." For that reason, the meaninglessness of BG is to be accounted for in the same way as the meaninglessness of:

(BG#) "The beautiful is blurga derba gurb the good."

The problem with BG# is that "blurga derba gurb" doesn't mean anything. It may be that its meaninglessness is *reflected* in its syntax. It may be that *because* "blurga derba gurb" doesn't mean anything, BG#'s syntax is off. But, if so, its syntactic shortcomings are to be explained in terms of its lack of meaning, not *vice versa*.

The same holds of BG. The reason that BG doesn't mean anything is that "more identical than" doesn't mean anything. BG's syntactic shortcomings are to be explained in terms of its lack of meaning, and not *vice versa.*

This is easily verified. Given any relation *R*, if "more identical than" denoted *R*, BG would be meaningful. If, for example, it denoted the relation that is in fact picked out by the expression "a better dancer than," BG would

say that the good is a better dancer than the beautiful; and it would thus have the same meaning as:

> (BG*) "The property of goodness is a better dancer than the property of being beautiful."

BG* isn't meaningless; it's false. The property of goodness can't dance; neither can any other property, including the property of being beautiful. Therefore, the former property isn't a *better* dancer than the latter. Nothing false is meaningless, since to be false is to bear a false meaning. So BG* is meaningful.

Incidentally, according to Gilbert Ryle (1900–1976), a career-long Wittgenstein-hardliner, it cannot meaningfully be said of properties that they can, or cannot, dance.[70] It can be said of a human being that he can, or cannot, dance. But nothing that can meaningfully be said of non-properties, such as human beings, can be meaningfully said of properties themselves.

Ryle's position is false. There are many differences between people and properties. One of them is that people can dance, whereas properties cannot. Ryle's position is also self-refuting. In saying that it cannot be said that properties cannot dance, Ryle is saying exactly what it is that, according to his theory, cannot be said.

Ryle is confusing absurdity with meaninglessness. The sentence "triangles have four sides" makes an absurd, but meaningful, statement. Given that triangles have three sides, not four, it's false. Given that it's false, it's meaningful. "Properties can dance" is absurd, and therefore false, and therefore meaningful.

There are two kinds of "nonsense." A sentence can be nonsense by failing to have *any* meaning. (BG# is nonsense in this sense.) And a sentence can be nonsense by virtue of having of an absurd meaning. ("Properties can dance" is nonsense in *this* sense.) Ryle doesn't distinguish between these two kinds of nonsense. Neither does Wittgenstein. And Wittgenstein's position, like Ryle's, is self-refuting—and for much the same reason. If Wittgenstein is right to say that all philosophical assertions are ungrammatical nonsense, that very assertion is ungrammatical nonsense and therefore isn't true.[71]

Given how implausible and illogical it is, why on Earth did Wittgenstein hold that philosophical statements are always ungrammatical nonsense? The answer, I believe, is that Wittgenstein thought this view to be the distillation

of Frege's groundbreaking philosophical successes. Frege showed that reparsing sentences sometimes solves philosophical problems. By reparsing:

(NS) "Nothing is a square circle,"

we show that it doesn't attribute the property of being a square circle to some non-entity, thereby dissolving an age-old riddle. Wittgenstein seems to have inferred from these successes of Frege's that *all* philosophical problems are to be solved by reparsing sentences.

This inference is fallacious. Given only that some philosophical problems are to be solved by reparsing sentences, it doesn't follow that they are all to be solved in that way. And we've seen that, indeed, many of them are not to be solved in that way.

Also, there were no cases where Frege's reparsing of a problematic sentence showed it to be ill-formed. For example, in reparsing NS, Frege showed that its logical syntax differed from its apparent syntax. He didn't show that there was anything *wrong* with its logical syntax.

To be sure, philosophers often produce deeply absurd statements. For example, according to pre-Fregean philosophers, NS entails that:

(NS_1) There exists some non-entity that is a square circle.

If NS_1 is right:

(NS_2) There exists some entity that does not exist that is a square and a circle.

NS_2 is self-contradictory and therefore absurd. But it's obviously meaningful. If it weren't, it wouldn't contradict itself. (For a sentence to contradict itself is for it to bear two opposed meanings. "Smith is a lawyer and Smith is not a lawyer" because the one ***conjunct***[72] contradicts the other.) But whereas this sentence is explicitly self-contradictory, the self-contradictory sentences that people actually utter are usually only implicitly so.

But supposing for argument's sake that philosophical statements are meaningless, it isn't *because* they're ungrammatical that they're meaningless.

Second, they're not meaningless. The statement that they're meaningless is itself a philosophical statement. That statement is therefore false if it's true. Therefore it is false. (Any statement that entails its own negation is false.)

4.3 The Tractarian roots of Logical Positivism

Although one of the TLP's contentions is that all philosophical statements are ungrammatical nonsense, this is not its main contention. The main contention of the TLP is that:

> (CT) A sentence is meaningful if, and only if, it is either a ***tautology*** or an ***observation report.***

A "tautology" is a definitional truth (e.g., "fathers are male," "there are three feet in a yard").

An "observation report" is a statement that reports what one's senses have told one (e.g., "I am now seeing a dog," "there is chocolate syrup (or, in any case, a brown discoloration of some kind or other) on Smith's ice-cream," "I can see your house from here").

CT can be broken down into two claims:

> (1) All meaningful non-empirical statements are *tautologies,*

and:

> (2) All meaningful non-tautologous statements are *observation-reports.*

(1) entails that non-empirical disciplines (e.g., philosophy, mathematics) consist of statements that say nothing about anything. (2) entails that anything non-tautologous that cannot be known *strictly* on the basis of what one's senses tell one is meaningless.

Wittgenstein's claim that all philosophical statements are ungrammatical nonsense is a ***corollary*** of CT. If they're meaningful, philosophical statements, unlike tautologies, are non-trivial. Consider the statement that:

(KC) "Knowing a truth doesn't necessarily involve one's being affected by the state of affairs described by that truth."

KC isn't empirical and neither is its negation. KC isn't a tautology and neither is its negation. But KC *is* a philosophical statement and so is its

negation. If CT is correct, it immediately follows that both KC and its negation are ungrammatical nonsense. Since philosophical assertions are *never* tautologous, CT entails, as Wittgenstein knew, that philosophical assertions are categorically meaningless.

The position that (1) and (2) are both correct is known as *logical positivism* (LP). During the decade or so following the publication of the TLP, and largely because of it, LP was very popular.[75] But (1) and (2) are false. Let us now say why.

4.4 Verificationism and falsificationism

(2) is identical with a doctrine known as ***verificationism.*** According to verificationism a non-tautologous statement is meaningful ***iff*** it's capable of being verified strictly on the basis of sensory observation.[76]

Verificationism is false. The statement "all metal expands when heated" is meaningful. But it cannot be conclusively verified (i.e., shown to be true) strictly on the basis of observation. No matter how many metal objects you find to expand when heated, it's a possibility that some metal object that you haven't yet considered will fail to do so.[77]

Even though it cannot be verified, "all metal expands when heated" can be *falsified* (i.e., shown to be false) strictly on the basis of observation. Advocates of LP saw this and, having for this reason rejected verificationism, they accepted a doctrine known as ***falsificationism.*** According to falsificationism, a non-tautologous statement is meaningful if and only if it's capable of being falsified strictly on the basis of observation.***[78]***

Falsificationism is false. Though obviously meaningful, the statement "there exists a gold ball that weighs exactly 27.13654 lbs." cannot be conclusively falsified, since no matter how many gold balls you consider it is possible that some gold ball that you haven't considered has that weight.

Falsificationism is really a version of verificationism. According to verificationism, *S* is meaningful if verifiable. According to falsificationism, *S* is meaningful if *S*'s negation is verifiable. So given that verificationism fails, it's no surprise that falsificationism does as well.

In light of the failure of falsificationism, advocates of LP decided to soften their views about meaningfulness one more time. This time, they said that a non-tautologous statement is meaningful if possible observations can ***confirm*** it. (*P* confirms *Q* if, other things being equal, *Q* is more likely to be true if *P* is true than if not-*P* is true. Other things being equal, Smith is more likely to be wealthy if he wears fancy clothes than if he doesn't wear fancy clothes. Thus, "Smith wears fancy clothes" confirms "Smith is wealthy.") We'll refer to this view as "confirmationalism."

Confirmationalism is equivalent with the position that all meaningful non-tautologous statements are ***empirical.*** An empirical statement is one that, if true, expresses a truth that must be known through observation and that, if false, is the negation of a true empirical statement. "There are trees in Santa Barbara, CA" is a true empirical statement, and "there are no trees in Santa

Barbara, CA" is a false one. The negation of an empirical statement is an empirical statement. This is because, if it's an empirical question whether or not *S* is true, then it's an empirical question whether or not *S* is false, and therefore whether or not not-*S* is true. (As we will see, there are very few truths, if any, that can be known *strictly* through observation. A statement is empirical if the truth or falsity of it is to be decided on grounds that are at least *partly* observational.)

Henceforth when we refer to "logical positivism" (LP), we will refer to the position that *a statement S is meaningful if and only if (i) S is a tautology (e.g., "sisters are female siblings") or (ii) S is an empirical statement (e.g., "there are trees in Santa Barbara").*[79]

4.5 Logical Positivism Evaluated

Consider the statement:

(1) “Triangles are three sided figures.”

(1) is plausibly seen as being true by convention. The same is true of:

(2) “Pentagons have five sides.”

For argument’s sake, we’ll grant that, indeed, (1) and (2) are true by convention—that they’re definitional truths.[80]

Given (1) and (2), it follows that:

(3) If *x* is the number of sides of an arbitrary triangle and *y* is the number of sides of an arbitrary pentagon, then *w* is an even prime iff *w* is one less than *x* and three less than *y*.

But (3) clearly isn’t a conventional truth. Though it follows from conventions, (3) is not itself a convention. (3) is thus a non-tautologous, non-empirical truth. It follows that LP is false.

It’s possible to have conventions that are ***inconsistent*** with one another. (Two statements are inconsistent with each other if it is logically impossible for them both to be true. Thus, “Smith has three cars” is inconsistent with “Smith has no cars.”) If I stipulate that “*x*” unambiguously refers to the number two, and I also stipulate that “*x*” refers to the number of sides of a triangle, my definitions are inconsistent with one another. The internal consistency of conventions is not itself a matter of convention. It isn’t an empirical fact that the conventions just described are inconsistent with each other. To say that *P* is inconsistent with *Q* is to say that *P must* be false if *Q* is true. Observation cannot tell you what *must* be the case; it can tell you, at most, what *is* the case. So it cannot tell you that two linguistic conventions are inconsistent with each other.

Since observation cannot tell you whether or not two statements are inconsistent with each other, it cannot tell you whether or not one statement is a necessary consequence of some other. This is because for *Q* to be a necessary consequence of *P* is for *P* to be inconsistent with the ***negation*** of

Q. (The negation of "John has an even, positive number of cars" is "it is not the case that John has an even, positive number of cars" or, more simply, "John does not have an even, positive numbers" In general, the negation of a *Q* is not-*Q*. "John has an even, positive number of cars" entails "John has at least one car" because the second sentence cannot be false if the first is true, i.e. the negation of the second sentence (namely, "it is not the case that John has at least one car" or, more simply, "John has no cars") cannot possibly be true if John has an even, positive numbers.)

When evaluating LP, one must be careful to distinguish sentences from their meanings. Two different sentences can have the same meaning (e.g., “snow is white” and “schnee ist weiss”). The meaning of a sentence is a ***proposition.*** A proposition is what is affirmed by an utterance or a sentence in the indicative mood. (So an example of a proposition is the thing that is affirmed by an utterance of “snow is white.”) Propositions are not sentences. The thing meant by “snow is white” is not itself a sentence. (“Schnee is weiss” is not the same sentence as “snow is white," but they both affirm the same proposition.) Some sentences express propositions that are logically true. A proposition is logically true if the laws of logic prohibit its negation from being true. The proposition meant by:

> (4) “If a given thing is round, then that thing is not a square,”

is logically true, since the laws of logic don’t allow round things to be square. For this reason, some authors would describe (4) itself, the sentence, as logically true. But what they mean is that *given* what it is that it means, (4) must be true. So what they are *in fact* describing as logically true is the proposition that (4) couldn’t be false, given what it is that it means.

Thus, it is always propositions, and never sentences, that are logically true. But no *proposition* is true by convention. It’s up to us what our symbols mean. But it’s not up to us whether those meanings are correct. It’s up to us what it is that “the moon is not made of cheese” means. But it isn’t up to us whether that meaning is correct. LP identifies logical truth with conventional truth: truths of logic are sentences that are true by convention. But that’s false. Logical truths are never sentences; they’re always propositions, and propositions are never conventionally true.

Incidentally, (ii) collapses into (i). Linguistic conventions are known

empirically. It can be known only through observation that “triangles have three sides” is true. That sentence could mean anything.[81] It could mean that penguins are smarter than humans; and it’s only because you’ve had the requisite sense-perceptions that you know it not to mean this.

It should be pointed out that logical positivists were unanimous in denying the existence of propositions and of meanings generally. Logical positivists didn’t accept the view that for a lecture to concern triangles is for the meaning of that lecture to concern triangles. This is why Rudolph Carnap[82] (1890–1970), an LP-hardliner for many years, said that for a lecture to concern triangles is for the word “triangle” to occur in the lecture.

Carnap’s view is false. Many a lecture that doesn’t contain the word “triangle” concerns triangles. (Think of all the mathematics lectures given in Japanese, Swedish, and Arabic. How often does the word “triangle” occur in them?) And a lecture that contains the word “triangle” isn’t necessarily about triangles. Somebody giving a lecture on linguistics may use the word "triangle" to illustrate some point about phonetics; but in so doing, that person isn’t talking about triangles. They’re talking, not about triangles, but about the *word* “triangle.”

The distinction between the word “triangle” and the corresponding meaning is one that Carnap couldn’t countenance without ceasing to be a logical positivist. The very essence of logical positivism is the denial of meaning. For argument’s sake, suppose there were to exist objects that are the meanings of sentences. (Following convention, we’ll refer to these things as “propositions.”) Given the existence of propositions, whether a given sentence is meaningful is *not* to be explained in terms of its being either tautologous or confirmable. Rather, a sentence’s being tautologous or confirmable is to be explained in terms of its bearing a proposition of a certain kind. A tautologous sentence would be one that had a logically correct proposition for its meaning, and a confirmable sentence would be one that had a confirmable proposition for its meaning. But if there are logically true propositions, then some truths are *ipso facto* not empirical.

We can use words in any way that we like. We can use the words “if Smith has three boats, then Smith has more than one boat” to mean that 1 + 1 = 3. But whatever meaning we end up assigning to those words, it’s not up to us whether that meaning is correct. And if that meaning is of a logical nature, as is actually the case, there is *ipso facto* non-empirical truth, an immediate consequence being that LP is wrong.

Carnap's attempt to do away with meanings consisted his saying (though he did not himself put it this starkly) that the meaning of the word "snow" was that very word and, in general, that words are their own meanings.[83] Given how brazenly wrong this view is, Carnap's attempt to do away with meanings never had many takers.

A much better received attempt to do away with meanings is to be found in a doctrine known as ***conceptual role semantics*** (CRS). According to CRS, two sentences have the same meaning if, and only if, they are *used* in the same way. So "hace mucho calor" is the Spanish translation of "it's hot" *not because* those sentences share a meaning—meanings don't exist, according to CRS—but because the one sentence is used in the same way as the other.[84]

CRS seems to coincide with the Wittgenstein thesis that "meaning is use." It is thus no surprise that, when scrutinized, CRS proves unfeasible. So far as she isn't misspeaking or randomly barking out noises, anyone who says "it's hot out" or "hace mucho calor," or any other sentence, does so because she knows that existing semantic rules assign a certain proposition to those words and she wishes to put that proposition into words.

One immediate consequence of this is that Carnap's position is false. Another immediate consequence is that what it is for those two sentences to have the same meaning is *not* for them to be "used in the same way."

Also, given any natural language, there are infinitely many sentences belonging to it that have never been used. The thesis that sentences have the same meaning iff they're used in the same way has the absurd consequence that any two sentences that haven't been used before have the same meaning. (If two sentences aren't used at all, they aren't used differently, and are therefore used in the same way – if only in a vacuous sense.)

4.6 Logical Positivism Evaluated (continued)

Let's resume our discussion of Logical Positivism (LP). We've seen that, contrary to what LP says, there are non-empirical truths that have nothing to do with anyone's linguistic practices. We'll now see that, contrary to what LP says, there are facts about the spatiotemporal world that cannot possibly be known strictly on the basis of sense-perception.

Let NT be the body of assertions jointly constituting Newton's physics.[85] There is no denying that NT is meaningful. But *by itself* NT doesn't make any predictions or otherwise have observable consequences. Physical laws are expressed by ***conditional*** assertions—that is, by statements of the form ‹if *P*, then *Q*.›[86] NT doesn't say anything about how this or that physical object will behave. It says how a given object will behave *if* certain conditions are met. For example, NT says an object will behave *if* it has a certain mass and is within a certain distance of another body having a certain mass.[87] But NT itself obviously doesn't say that this or that object has this or that mass or is within this or that distance of this or that other specific body. Thus, *taken by itself,* NT isn't confirmable. What *is* confirmable isn't NT, but NT *plus* statements describing specific matters of fact. In other words, what is confirmable isn't NT, but rather some statement of the form ‹given such and such, NT makes it likely that thus and such.›

But NT is obviously meaningful. This is a problem for LP. To deal with it, advocates of that doctrine proposed that (ii) be replaced with the position that (ii*) a statement S_1 is meaningful if there is some statement S_2, such that *given* S_2, S_1 is confirmable (i.e., capable of being supported by observation).

By this standard, "the nothing nothings" qualifies as meaningful and so does every other nonsense sentence one can think of. Given the statement "if grass is green, then the nothing nothings," anything that confirms "grass is green" confirms "the nothing nothings," and "the nothing nothings" thus qualifies as meaningful.

LP replaced (ii*) with other, similar proposals. But they all failed for reasons similar to the one just discussed.[88]

4.6.1 An incoherence internal to verificationism and falsificationism

In this section, we'll identify an incoherence that lies at the heart of both verificationism and falsificationism. But before we can identify that incoherence, we must make a preliminary point concerning what it is that verificationism and falsificationism allege. (For the sake of brevity, we'll focus on verificationism, leaving unstated the obvious corollaries concerning falsificationism.)

Verificationism is not the doctrine that in order for a statement to be meaningful it must *in fact* have been verified. There are several reasons why such a doctrine would be too absurd to have so much as a veneer of plausibility. First of all, it would have the false consequence that false statements couldn't be meaningful. (Since meaningless statements are neither true nor false, false statements are always meaningful.) Secondly, such a doctrine would have the consequence that statements that haven't been verified are categorically meaningless---even if they *could* be verified. So, supposing that no one has looked inside the cupboard, it would have the consequence that, since it hasn't been verified, the sentence "there is bread in the cupboard" is meaningless. Thus, such a doctrine would have the consequence that statements are meaningless if we cannot *yet* verify them, even though we *could* verify them were circumstances to change. So such a doctrine would entail that, if we didn't have the technological wherewithal to open the cupboard, the statement "there is bread in the cupboard" would be meaningless *even though* (a) it obviously does have a meaning and (b) we might soon have the requisite technology. Finally, such a doctrine would entail that hitherto meaningless statement would *become* meaningful upon being verified. So "there is bread in the cupboard" would be meaningless *until* it was verified; and, upon being verified, it would be meaningful.

But statements cannot become meaningful by being verified. Statements aren't in a position to be verified *unless* they are *already* meaningful. Because it has no meaning, "yarblek truflo blurgo" cannot be verified. And it's only because "there is bread in the cupboard" does *have* a meaning that it can be verified.

For these reasons, verificationism, if it is to have any plausibility at all, cannot possibly be the doctrine that a statement is meaningful iff it *has* been verified, and it must instead be the doctrine that a statement is meaningful iff it *could* be verified, at least in principle. So verificationism must be the

doctrine that a statement *S* is meaningful iff there are *possible* observations that definitively establish its truth. In other words, *S* is meaningful iff, setting aside the limitations on our powers of observation set by our current circumstances, there are possible observations $O_1 \ldots O_n$ such that, if those observation were had, *S* would *ipso facto* have been proven correct.

Bearing these points in mind, consider the following passage:

> (*) There are *possible* observations that, were they actual, *would* verify statement *S*. To put this in a way that doesn't suggest, absurdly, that there exist non-existent objects: there are observation-*types* that are not instantiated but whose instances, supposing that the aforementioned types *were* instantiated, *would* verify *S*.

We've seen, unless (*) is meaningful, verificationism collapses into the brazenly absurd doctrine that a statement is meaningful only if it *has* been verified---only if, in other words, *actual* observations have proven it correct.

But, supposing that verificationism is correct, how is (*) to be verified? Not through observation. For *ex hypothesi* the relevant observation-types have not, as of yet, been instantiated. So there is no direct, experiential evidence that those types *could* be instantiated. Thus, (*) is false unless there exist entities for which there can be no direct experiential or observational evidence. Therefore, according to verificationism, (*) is meaningless. No meaningless statement is true. Verificationism is false unless (*) is true. Therefore, verificationism is false.

Also, if verificationism is right, claims about what is possible but non-actual are meaningless. What is non-actual cannot be observed and claims about it thus cannot be verified. So (*) is meaningless if verificationism is correct, and verificationism is therefore false.

Falsificationism ("*S* is meaningful iff not-*S* is verifiable") is false for the same reasons *mutatis mutandis* as verificationism.

4.7 Logical Positivism self-defeating

For the reasons just given, it soon became clear that LP was unsalvageable, and soon everybody jumped ship. In fact, it was some of LP's most staunch proponents who first made it clear what LP's shortcomings were.

We'll see this in this chapter when we discuss the brilliant criticisms of LP put forth by Carl Hempel (1905–1997), who was one of LP's first and most ardent advocates. But erstwhile advocates of LP tended not to see the incoherence that lies at the center of that doctrine. When saying why they rejected LP, they usually cited narrow, technical problems of the sort just discussed. What they didn't do, but what we're about to do, is to say why LP is at its very core a broken and illogical doctrine.[89]

Anything that is true or false is meaningful. Truth implies meaningfulness and so does falsity. Thus, LP is meaningful if it's correct. LP says that any meaningful statement is either a tautology or is empirical. So if LP is correct, it is itself either a tautology or it is empirical.

It isn't a matter of convention that the expression "meaningful sentence" is interchangeable with the expression "sentence that is either a tautology or is empirical." Therefore, LP isn't a tautology.

Since it isn't a tautology, LP is an empirical truth if it isn't false. (A similar argument: LP is a theory. It is a theory concerning the conditions sentences must satisfy to be meaningful. Tautologies aren't theories. Tautologies are conventions. Conventions aren't theories. Therefore, LP isn't a tautology.)

But LP isn't an empirical truth. Any attempt to provide an observational basis for any statement presupposes the meaningfulness of that statement and thus presupposes an answer to the question "what conditions must a statement fulfill to be meaningful?" For this reason, the question "what conditions must a statement satisfy if it is to be meaningful?" isn't empirical in nature. It follows that one cannot coherently attempt to find *empirical* grounds for accepting LP, since any attempt to do so *itself* presupposes the meaningfulness of LP. Thus, LP isn't an empirical theory.

Thus, LP is neither an empirical truth nor a tautology. It is thus a counterexample to itself and is therefore false.

Interestingly, in the TLP, Wittgenstein seems to be aware that one cannot coherently say that there can be non-tautologous meaningful statements. Not a single one of the assertions in the TLP is empirical; and not a single one of

them is a tautology. This means that, if the TLP's main thesis is correct, the TLP is nonsense.

Wittgenstein acknowledges this. For he ends his book by saying that every point that he has made is meaningless and, consequently, that not a single one of them is true. But, Wittgenstein says, those enable one to *see* the truth, even though they are not themselves true. These non-truths, Wittgenstein says, are like ladders that help one get to the truth. (I would have used a different metaphor, e.g. "much as telescopes help us see the stars even though they themselves obviously aren't stars, so the points Wittgenstein makes help us see certain truths, even though they are not themselves truths.") Here are the very last words of the TLP:

> My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them—as steps—to climb up beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.) He must transcend these propositions, and then he will see the world aright. What we cannot speak about we must pass over in silence.[90]

So far as they aren't trivial, these magisterial words are false. To understand something *is* to see its meaning. Therefore, anything that is understood has a meaning. So Wittgenstein's words, if understood, have a meaning, and Wittgenstein is therefore contradicting himself in saying that those who understand his words will see that they're meaningless. Wittgenstein's awe-inspiring injunction that we pass over in silence what we can't speak about involves a similar solecism. To remain silent about something *is* to pass over it in silence. So Wittgenstein is asking, emptily, that we not say anything about what we can't possibly say anything about.

4.7.1 Empiricism self-refuting

These points are easily extended to show that ***empiricism*** is false if it's true and, therefore, that it's false.

Empiricism isn't the claim that:

> (1) Whatever we know *now,* we learned it through sense-perception—but it's possible that at some time in the future we'll acquire knowledge in some other way.

Empiricism is the doctrine that everything that

> 2) It's inherent in the nature of knowledge that all knowledge be strictly observation-based.

But it cannot be known through observation that (2) is correct. According to (2), ‹*x* is knowledge› is inconsistent with ‹*x* isn't known through observation.› But, as we noted on page 24, observation cannot tell you whether one statement is consistent with some other statement. Thus, any body of observational data is consistent with the assumption that empiricism is false.

This means that there cannot be strictly observational grounds for believing empiricism correct. Thus, so far as empiricism is correct, there are no grounds for believing it correct; and so far as there are such grounds, empiricism is false. Thus, the likelihood that empiricism is correct is inversely proportional to the degree of probability that the information at our disposal confers on it. And this means that if it's a certainty that empiricism is correct, it's a certainty that it's false. Therefore, empiricism, if true, is false; therefore it's false.

4.7.2 A Famous Argument of Russell's

A famous argument given by Bertrand Russell (1872-1970) ends with a sentence very similar to the last one. (This was deliberate.) Since the argument in question is to the effect that "naïve realism" is false, naïve realism being a doctrine that falls within the scope of the present chapter (though the reasons for this are not easily stated, and therefore won't be), and we will therefore critique it, after first quoting it *verbatim*:

> We all start from naïve realism, i.e., the doctrine that things are what they seem. We think that grass is green, that stones are hard, and that snow is cold. But physics assures us that the greenness of grass, the hardness of stones, and coldness of snow are not the greenness, hardness, and coldness that we know in our experience, but something very different. The observer, when he seems to himself to be observing a stone, is really, if physics is to be believed, observing the effects of the stone upon himself. Thus, science seems to be at war upon itself. When it most means to be objective, it finds itself plunged into subjectivity against its will. Naïve realism leads to physics; and physics, if true, shows that naïve realism is false. Therefore, naïve realism, if true, is false; therefore it's false.[91]

Though eloquently stated, this argument consists of spurious reasoning for a false conclusion. In observing the paper-weight on my desk, I'm not observing some effect of the stone upon myself: I'm observing the stone itself.[92] To be sure, my observing the stone is *itself* an effect of some event involving the stone. (Light bounces off the stone and, in due course, strikes my retinas, precipitating various physiological and psychological responses, among them the aforementioned sense-perception.) But that doesn't mean that *what* I'm seeing, in having that sense-perception, is some effect that the stone had on me; and unless we're perpetually hallucinating, in which case we're never observing anything external, we very obviously do observe external objects.

This brings us to the second problem with Russell's argument. Russell says that the greenness, hardness, and coldness of daily observation are not identical with their counterparts in physics. This is false. Physics has a lot to

say about greenness, hardness, and coldness that commonsense does not. But that's very different from saying that the coldness we feel, the greenness we see, etc., *aren't* the greenness, coldness, and hardness of physics. Physics tells us *what it is* for something to have the properties we know them to have through sight, touch, etc. When you grab an object and feel that it's cold, you don't feel or otherwise sense-perceive the micro-events *in virtue of which* it is cold. (More precisely, when you grab an object and feel that it's cold, its coldness is not thereby *presented to you* as consisting of the micro-events of which it does, in fact, consist.) It is the job of the theoretical physicist to tell you about these micro-events. This means that the coldness studied by the physicist *is* identical with the coldness that you feel. Physics tells us that many of our pre-theoretic beliefs as to what *that coldness is* are wrong; and in order to do that, it has to study the coldness that those pre-theoretic beliefs concern. So Russell's argument, despite Einstein's high regard for it, is a failure.[93]

4.7.3 The objects of perception distinct from their causes

A moment ago I said that "[W]hen you grab an object and feel that it's cold, you don't feel or otherwise sense-perceive the micro-events *in virtue of which* it is cold." Many are likely to object to this: All of our sensations are the results of micro-events, it will be said, and there is thus is *a* sense in which we perceive such events.

This line of thought involves the spurious belief that the cause of a perception is identical with its object. When I see your face, the immediate causes of my perception are various disturbances of retinas. But I don't perceive those disturbances. In order for an experience *x* to be a perception, as opposed to a mere effect, of some event *y*, it is necessary that, by virtue of having *x*, one have some knowledge of what *y* is. By virtue of seeing your face, I am aware of the structure of your face. By virtue of feeling a surface, I am aware of the texture of that surface. Visual and tactile perceptions resemble their objects. My visual perception of your face has the same structure as your face, that being why, by virtue of having that perception, I am aware, within limits, of the structure of your face.

To be sure, the object of a given perception is *among* the causes of that perception. I cannot see your face unless I am in some way affected by it. (Light rays must bounce off of your face, then strike certain surfaces of mine, and so on.) But the object of a perception is always a distal, never a proximal, cause of that perception.

Not all perceptions bear any (obvious) structural resemblance to their objects. For example, if I hear a loud noise, I am not, by virtue of that fact, aware of the structure of the object of my perception. Nonetheless, I am, by virtue of having that experience, dimly aware of facts about the external world. I am aware that there has been an abrupt change of some kind. I am dimly aware of the location and distance, relative to myself, of that noise. Similar points hold with respect to olfactory and gustatory experiences. The unpleasant experience of smelling rotten food doesn't bear any resemblance to its object. But, by virtue of having that experience, I am aware of some irregularity in the world that I wouldn't otherwise be aware of, and I am aware of at least some basic facts about it.

But, supposing that I am having some sensation caused by various micro-events, I am not, by virtue of having that sensation, aware of those micro-events at all. I couldn't possibly hear a loud noise or smell a foul odor

without thereby knowing of the existence of *some* object in the external world of which I'd otherwise be ignorant. But people can obviously have sensations that are caused by micro-events without knowing of the existence of those events. In fact, most people, when having sensations, haven't the faintest idea that their experiences are by-products of sub-atomic activity. So merely having a sensation that is caused by such and such micro-events isn't enough for perceiving those events.

Russell says that the greenness and the hardness studied by the physicist are distinct from the greenness and the hardness known to the non-physicist. That is absurd. The leaf that the botanist looks at through a microscope is obviously identical with the leaf looked at with the naked eye by the non-botanist. The leaf is a complex entity. The botanist knows more about it than the non-botanist. But that doesn't mean that the leaf studied by the botanist is distinct from the leaf observed by the non-botanist. A given leaf's color is a complex entity; and the physicist knows more about that entity than the non-physicist. But it obviously doesn't follow that the green studied by the physicist is distinct from the green known to the non-physicist. The same is true of the leaf's temperature and mass and, indeed, of any feature of anything that is both known to the layperson and scrutinized by the scientist.

4.7.4 The empiricism-unfriendliness of the concept of confirmation

Confirmationalism, the doctrine that a non-tautology is meaningful iff confirmable, is a form of empiricism. But we'll now see that the concept of confirmation is an incoherent one *unless* it's granted that there is non-empirical knowledge and, therefore, that confirmationalism is incoherent.

An argument put forth by Nelson Goodman (1954) makes this clear:

> (GA)[94] Let's say that an object is "grue" if it's green and examined before Jan. 1, 2010, or it's blue and examined anytime thereafter. All green objects examined before Jan. 1, 2010, are grue. So, supposing that we've examined ten million emeralds before Jan. 1, 2010, and found them all to be green, we've also found them to be grue. Presumably, the fact that they've all been green warrants the inference that they'll be green after Jan. 1, 2010. But, so far as that data entitles us to infer that they'll be green, it also entitles us to infer that they'll be grue and, therefore, blue after Jan.1, 2010.
>
> This line of thought is easily extended to show that anything can confirm anything.[95] Let phi, psi, and chi be three properties such that (i) a thing is phi if it's examined by a human being who knows and therefore truly believes that, at that time, no human being can fly; (ii) a thing is psi if examined by a human who knows that, at that time, all human beings can fly; and (iii) a thing is chi if it's examined before Jan. 1, 2010, and known to have phi or it's examined after that time and is known to have psi. Since everything ever examined as of the present time (May 24, 2009) has had phi, it's also had chi. So given that, thus far, no human has been able to fly, we're no less entitled to infer that in 2015 they'll all be able to fly than we are entitled to infer that, at that same time, none of them will.

GA can be taken to show *either* that no inductive inference is better than any other *or* that, since some inductive inferences clearly are better than others, GA must involve an error of some kind. Supposing that the second interpretation is the right one, it's easy to identify the problem with GA.

Contrary to what that argument tacitly assumes, induction does not have strictly observational basis. From a strictly *observational* standpoint, it's no less correct to describe an emerald examined in 2009 as "grue" than it is to describe it as "green."

Given any body of data, there are different, but equally *observationally legitimate* ways of describing it. This means that, if any inductions are better than any others, we must have legitimate but at least partly *non*observational grounds for believing that, when making inductive inferences, certain properties (e.g., *green*) are relevant and others (e.g., *grue*) are not. (These grounds are identified in Chapters 17.)

4.8 The brokenness of the concept of tautological truth

In this section we'll see that it is utterances of sentences, not sentences per se, that are tautologies. For reasons that will become clear, this entails that, contrary to what LP alleges, non-empirical truth cannot be identified with conventional truth.

Whether a given utterance is tautologous very much depends on the manner in which the person hearing that utterance learned the meanings of the expressions composing it. A story may help make this clear. You don't know to what length the word "yard" refers, and you ask your friend Smith to give you this information. In response, he points to some object *L* and says: "the length of *that* object is one yard." *L* is *in fact* three-feet long. But you can't tell this just from looking at it. You can tell roughly, but not exactly, how long *L* is. You don't bother to measure *L*. This all happens on Monday.

The next day, you see some object *M*. You measure it and find that its length is three feet. You tell your (still present) friend Smith that *M* is exactly three-feet long. You know, of course, that *M*'s length is more or less comparable to *L*'s—that neither length is, for example, ten times as great as the other. But you don't have precise knowledge of their comparative lengths; you don't know, for example, whether *L*'s length is within six inches of *M*'s. Because you have a passion for knowing the comparative lengths of objects, you find this upsetting, and you tell Smith that you wish you knew how *L*'s length compared to *M*'s. Smith says: "I don't know why you're upset. You've measured *M* and found that it is three-feet long. Since, as you know, *L* is a yard long, it's patently obvious what *M*'s length is." You don't quite know what he means, and you tell him this. Somewhat irritated, he says:

(i) "There are three feet in a yard."

Under these circumstances, (i) is *not* trivial and, therefore, is *not* a tautology. It *would* be tautologous if you had learned the meaning of the word "yard" by being told that "yards are lengths of three feet." But this isn't how you learned that word's meaning. You were *shown* a yard-long object and told that the word "yard" refers to its length. Obviously that visual perception *did* apprise you of that object's length. But the *way* it described that length to you was different from the *way* that this same information would be conveyed to you by an utterance of: "the length of *that* object is one

yard."

And given the information embodied in your visual perceptions of *L* and *M*, it wouldn't be a trivial matter to know that the length described by the contents of your *L*-perceptions coincided with that described by your *M*-perceptions. Therefore, (i) would not, under those circumstances, express a tautology, at least not from your perspective. But it *would* express a tautology from the perspective of somebody to whom "yard" had initially been defined as "distance of three feet."

So even though it's standard practice among philosophers to describe *sentences* as "tautologies," this practice embodies a serious confusion. A given sentence may or may not be tautology, depending on the manner in which the auditor learned the meanings of its constituent expressions and depending, therefore, on the information *on the basis of which* the auditor knows those meanings.

It might be thought that, so far as (i) is non-trivial to you, it's only because you don't *really* understand it. This isn't true. To somebody who doesn't speak Albanian, a sentence of Albanian isn't trivial or non-trivial. It doesn't mean anything to such a person and is, from his perspective, just so much noise and is no more "trivial" or "non-trivial" than the sound of wind chimes. But, in our story, you *do* know what is meant by (i). It's not as though you're hearing a sentence of Albanian (or, if you happen to speak Albanian, a sentence of some language that you don't know). Therefore, it is only because you understand Smith's utterance of (i) that it is nontrivial for you. Therefore, tautologousness and non-tautologousness are properties, not of sentences, but of the information on the basis of which auditors figure out the meanings of sentences. And it's wrong to say that sentences *per se* are, or are not, tautologies.[96]

This story illustrates some deeply important facts that are in the philosophy of language. First, one knows the meanings of expressions descriptively. It is through sight and hearing (and other sensory modalities; e.g., touch) that you learn what words mean. Your perceptions apprise you of facts about the world by describing them to you—by apprising you of their colors, shapes, etc. Two very different descriptions can pick out some one thing (cf. "the third U.S. President" and "the President responsible for the Louisiana Purchase"). Therefore, the perceptually encoded descriptions *through* which one learns what two expressions mean may differ enormously, *even* if those expressions mean the same thing. A consequence is that what

utterances tell you is as much a function of the information through which you learn their meanings as it is of those meanings themselves. Thus, a given sentence may convey very different propositions to different people, all of whom know what it means, the reason being that those people access that meaning through different descriptions.

4.9 An alternative to the logical positivist conception of meaningfulness

The logical positivist's analysis of meaningfulness was a complete failure. I'd like to propose an alternative to it.

First of all, when asked to give examples of meaningless statements, the logical positivists tended to cite sentences that nobody ever uses; for example, "the nothing nothings,"[97] "the all is one," "the absolute is perfect." This is deeply suspicious: a theory that only takes on straw men can't be much of a theory.

And LP can't even prevail against these straw men. For, contrary to what its advocates said, the problem with these so-called statements is *not* that they're incapable of empirical corroboration. Consider the sentence:

(i) "The universe is a perfect unity."

Much loved by many a freshman narco-intellectual, this is a meaningless sentence if ever there was one; and it is just the sort of sentence that logical positivists had in mind.

But if it were said what exactly it means to describe something as a "perfect unity," (i) *would* be meaningful, as it would then be true or false. If, by a "perfect unity," one means an object that consists of events bearing a specified causal or logical relationship to one another, then (i) is either true or false, depending on the identity of that relationship. For example, if a "perfect unity" is an object such that, given any two non-simultaneous events composing it, there is a possible causal process connecting the first of those two events with the second, then (i) is true. (In contemporary physics, ‹*x* precedes *y*› is defined as: ‹there is a possible causal process, e.g., a light-signal, beginning with *x* and ending with *y*.›) On the other hand, if by a "perfect unity" one means an object such that if *x* and *y* are any two of its parts, the very idea of *x* existing in the absence of *y* is an incoherent one, then (i) is meaningful—and false. (One can coherently imagine a universe in which Cheney exists but Biden does not.)

In any case, the term "perfect unity" is clearly intended to refer to some sort of causal or logical integratedness, and once that mode of integration is pinpointed, (i) speedily becomes a true or false claim.

"Statements" such as "the universe is a perfect unity," "the nothing

nothings," and so on, aren't really statements at all. They're ***statement-forms***. (A statement-form is what results when one of the constants in a *bona fide*sentence is replaced with a variable. So "*x* is an even number" is a statement-form, since it is what results when the occurrence of "4" in "4 is an even number" is replaced with a variable. "4 is an even number" is an *instance* of the statement-form "*x* is an even number," and so is "8 is an even number." Unlike their instances, statement-forms are neither true nor false.) But (i) is obviously supposed to attribute some property to the universe. Since this property isn't identified, (i) contains an undefined term and therefore says nothing. Once that term is defined, a meaningful statement results.

It's true that (i) is neither tautologous nor confirmable. But that's a *consequence* of the real problem, viz., that "perfect unity" is undefined. (i) is comparable to "*x* is tall." The reason "*x* is tall" says nothing is that "*x*" is undefined—it hasn't been assigned a referent. And no sooner is a referent assigned to "*x*" than "*x* is tall" becomes meaningful. ("*x* is tall" comes to have the same meaning as "Bob Dole is tall" the moment Bob Dole is assigned to "*x*.")

Before "*x*" is assigned a referent, "*x* is tall" is neither confirmable nor tautologous. But that's only a symptom of the real problem, viz., that "*x*" doesn't have a referent. The same thing *mutatis mutandis* is true of (i).

Let's move onto the next phase of our argument. If *S* is a meaningful sentence, there is some object *x* and some property phi such that *S* says that *x* has phi. In other words, any given sentence is equivalent to one that has the form: ‹*x* has phi.› Let us now say why this is so.

Any non-compound sentence (i.e., any sentence that doesn't consist of other sentences) either says of some individual or item that it has some property, or it says that two or more individuals or items are interrelated in a certain way. Thus, "Smith is tall" says of some individual (Smith) that he has a certain property (tallness), and "Bob loves Sally" says that one individual (Bob) bears a certain relation (that of loving) with respect to some other individual (Sally).

"Smith is tall," "Jerry snores," and all other non-relational, non-compound sentences obviously have the form ‹*x* has phi.› The same is true of "Bob loves Sally," "Wilma detests Linda," and all other non-compound sentences that affirm the existence of relations between two or more objects. Let "*R*" be defined as follows: for any objects *x* and *y*, ‹ <*x*,*y*> has R› is true iff *x* loves *y*. (So for any objects *x* and *y*, the ordered pair <*x*,*y*> has *R* exactly if *x* loves *y*.)

Thus, “Bob loves Sally” is equivalent with “<Bob, Sally> has *R*,” which has the form ‹*x* has phi.›

A similar procedure can be performed on sentences (such as “Bob is standing in between Sally and Larry”) that affirm the existence of relations involving three or more objects. Let “*R**” be defined as follows: for any objects *x*, *y*, and *z*, ‹<*x*,*y*,*z*> has *R**› is true iff *x* is standing in between *y* and *z*. Thus, “Bob is standing in between Sally and Larry” is equivalent with “<Bob, Sally, Larry> has *R**,” which has the form ‹*x* has phi.› Other non-compound relational sentences are to be dealt with similarly.

What about compound sentences? Not a problem. Let “*K*” be defined as follows: for any sentences S_1 and S_2, ‹ <S_1, S_2> has *K*› is true iff the state of affairs described by S_1 is a consequence (of some kind or other) of the state of affairs described by S_2. Thus, “Smith broke his leg because he fell out of a tree” is equivalent with “<Smith broke his leg, Smith fell out of a tree.> has K,” which has the form ‹*x* has phi›. Other compound sentences are to be dealt with similarly.

Negative sentences are particularly easy to deal with. “Smith doesn’t smoke” is equivalent with “the proposition that Smith smokes is false,” which clearly has the form: ‹*x* has phi.› Other negative sentences are to be dealt with similarly.

There is only one kind of sentence that we haven’t yet considered, namely, ***quantified generalizations.*** A “quantified generalization” is any statement that says how many members one class of objects has in common with some other class of objects. Examples are: (a) “some person smokes,” (b) “no giraffes fly,” and (c) “all mice read Tolstoy.” (a) says that the class of people has at least one member in common with the class of smokers. (b) says that the class of giraffes has no members in common with the class of things that fly. And (c) says that the class of mice has no members in common with the class of things that don’t read Tolstoy.

Bearing this in mind, let “*E*” be defined as follows: for any properties *P* and *Q*, ‹ <*P*, *Q*> has *E*› is true iff the class of things having *P* has a least one member with the class of things having *Q*. Thus, (a) is equivalent with: “<the property of being a person, the property of being a smoker>, has *E*” which obviously has the form ‹*x* has phi.› Other quantified generalizations are to be dealt with similarly.

We have thus established that any given sentence *S* is equivalent to some sentence having the form ‹*x* has phi.› Given this fact, there is an obvious

answer to the question “what is it for a sentence to be meaningful?” A sentence is meaningful if it attributes some property to some object. A sentence *S* is meaningful if, for some object *x* and some property phi, *S* says that *x* has phi. It’s irrelevant whether it can be perceptually confirmed, let alone verified, that *x* has phi.

How could this theory be wrong? If a sentence attributes any property to any thing, it says something about something and is therefore meaningful. And if a sentence doesn’t attribute any property to anything, it doesn’t say anything about anything and is therefore meaningless.

5.0 The picture theory of meaning

One of the most interesting contentions put forth in the *Tractatus Logico-Philosophicus* (TLP) is the so-called "picture theory of meaning." In the TLP, Wittgenstein says that sentences are "pictures" of the facts they describe.

What does he mean? Maybe he means that sentences are picture-*like* in that they, like pictures, represent facts. But in that case what Wittgenstein is saying is completely and utterly trivial.

Thus, so far as what Wittgenstein is saying has substance, it isn't that sentences are *like* pictures of the facts they describe; and it must therefore be that they *are* such pictures.

But in that case, what Wittgenstein is saying is false. It is only relative to arbitrary conventions that "Smith punched Jones" describes the fact that Smith punched Jones. But it isn't relative to such conventions that a film or painting of Smith punching Jones describes that fact.

To be sure, there is a non-conventional *component* to sentential representation. Let *P* be the proposition meant by the sentence:

> (1) "Given that Socrates was a philosopher, it follows from the fact that Socrates was bald that there has been at least one bald philosopher."

The fact that (1) means *P* is not itself a convention. It is a logical *consequence* of our semantic conventions (e.g., that "Socrates" refers to Socrates, etc.). But for that very reason, there is a conventional component to that fact. And since graphic resemblance is a *non*-conventional method of representation, it follows that, so far as (1)'s relation to the fact it describes is conventional, that relation is fundamentally *not* like the relation borne by a picture of an event to the event itself. Thus, the picture-theory is false if taken literally and it is empty if taken non-literally.

But maybe there's some way of interpreting that theory that we've overlooked. To see whether this is so, let's consider Wittgenstein's own way of wording his own argument for it:

> At first sight a sentence—one set out on the printed page, for example—does not seem to be a picture of the reality with which it is concerned. But neither do written notes seem at first

> sight to be a picture of a piece of music, nor our phonetic notation (the alphabet) to be a picture of our speech. And yet these sign-languages prove to be pictures, even in the ordinary sense, of what they represent[98] . . . There is a general rule by means of which the musician can obtain the symphony from the score, and which makes it possible to derive the symphony from the groove on the gramophone record, and, using the first rule, to derive the score again. That is what constitutes the inner similarity between these things which seem to be constructed in such entirely different ways. And that rule is the law of projection which projects the symphony into the language of musical notation. It is the rule for translating this language into the language of gramophone records.

The idea seems to be that just as laws of projection coordinate the painting of the bowl of fruit with the bowl of fruit itself, so the ***semantic rules*** of a language coordinate its sentences with the realities they describe. (A semantic rule is what is expressed by any sentence saying what is meant by an expression of a language. So if, while pointing to individual *x*—who, we will stipulate, is Mick Jagger—I say “that guy’s name is Mick Jagger,” I’m stating a semantic rule, since what I’m saying is that “Mick Jagger” refers to *x*.)

But this analogy is a shallow one; and when it’s scrutinized, it becomes even apparent than before how *un*like pictures sentences are. If, because of some change in the environment, snow turned black, the English language would, without having to change, give us the resources to describe this change. We could describe it by saying “snow is black.” Thus, as far as the English language is concerned, snow can be any color. The semantic rules of English don’t say that snow *is* white. They say that, *if* snow is white, one can express that fact by saying “snow is white.” Thus, the semantic rules of English do not associate the sentence "snow is white" with the *fact* that snow is white; for those rules leave it open whether or not snow is white. Rather, they associate "snow is white" with a proposition such that, *if* it's true, it is a fact that snow is white.[99] Thus, sentences depict facts *by way of* having propositions for their meanings.

But this isn’t how photographs work. A photograph doesn’t go through the corresponding proposition. It goes straight to the fact (when there is one). No

picture goes through a proposition. Pictures, unlike sentences, go straight to the facts, if any there be, that they represent. This shows how fundamentally unlike pictures sentences are.

A related point is that pictures have structures that are radically different from those of any sentences. Sentences are ***digital*** structures.[100] They have a unique decomposition into a finite number of discrete parts. ("The cat is on the mat" decomposes into "cat," "mat," etc.) Pictures aren't like this. A picture of a cat on the mat doesn't have one, minimal unit of significance corresponding to the cat, another to the mat, etc. The part of the picture corresponding to the cat may also contain a part corresponding to the cat's ear and to the cat's leg, etc.

The fact that sentences, unlike graphic representations, are digital structures is a consequence of the fact that the former, unlike the latter, have a conventional component. The reason for this is a subtle one. But it's worth stating since it shows how deeply wrong the picture-theory is and since, in so doing, it reveals a lot about language.

Let D_1 be some random photograph of a person smiling. D_1 isn't a symbol of a language. But that could easily change. For this to happen, some convention would have to arise whereby it had a fixed a meaning—whereby it meant, for example, that people are sometimes happy, and the same thing *mutatis mutandis* happens in connection with each of several other photographs. So for example, there is some photograph D_2 of a person who is crying, and some convention is created whereby D_2 means that people are sometimes unhappy; and so on. Let L be the language defined by the totality of these conventions.

Even though D_1 is an image, it isn't functioning as an image so far it's functioning as an expression of L. The fact that D_1 is a picture of a smiling person may obviously make it easier for people to remember that, in L, D_1 means that people are sometimes happy. But it won't be *what it is* for D_1 to bear that meaning, or any other, in L. The character "0" is an unfilled hole and can thus be taken as graphic representation of emptiness. But so far as "0" is functioning in *that way*, it isn't functioning as a numeral that picks out zero. The same point *mutatis mutandis* holds of D_1.

Also, D_1 *doesn't* graphically represent the fact that people are sometimes happy; it graphically represents the tenuously related fact that, on some one

occasion, some one individual was happy, along with various other specific facts about that person's appearance that have nothing to do with anyone's being happy. So it isn't by virtue of graphically representing the fact that people are sometimes happy that, when functioning as a sentence of *L*, D_1 describes that fact.[101]

So far as D_1 is a sentence of a language, what it actually picks out is irrelevant; the various nuances of the smiling gentleman's face are irrelevant. In general, its internal structure is irrelevant. Considered as an expression of *L*, it *has* no internal structure. (It is what philosophers of language call a ***semantic primitive*.** A "semantic primitive," or "primitive symbol," is one that doesn't consist of other symbols and that, so far as it is an expression of a language, thus has no internal structure.) For exactly similar reasons, each of the other photographs composing *L* is, when considered as an expression of *L*, devoid of internal structure and thus, in the most extreme way possible, *not* like a graphic representation.

We must make one more point before we can close the argument. Let *N* be a photograph of a bolt of lightning, and suppose that *N* is the *L*-translation of the English expression "it is not the case that." So if *S* is the *L*-translation of "grass is green," *NS* is the *L*-translation of "grass is not green." (*NS* is formed by putting *N* to the left of *S*.) Even though *NS* consists of pictures, it is not itself a picture. Putting two pictures together isn't one picture; it's just two pictures that are next to each other. For exactly similar reasons, if conventions were created whereby the sentences of *L* could be ***conjoined, disjoined,***[102] or otherwise combined, the resulting compound sentences would not be pictures.

Let us take stock. Not a single one of the simple symbols belonging to *L* is a graphic representation of the fact that it depicts, and not a single one of the complex symbols belonging to *L* is a graphic representation of the fact depicted by any one of its components. In general, to the extent that a given thing is functioning as a linguistic expression, it is not functioning as a picture. Things that happen to be pictures cannot function as pictures so far as they are functioning as linguistic expressions or, therefore, as sentences. Thus, Wittgenstein's contention that sentences are pictures of the facts they describe is the antithesis of the truth. In addition to showing that Wittgenstein's picture-theory is false, this shows that *any* conventional assignment of meaning to *any* collection of symbols—any *language*, in other words—necessarily yields symbols that have a unique decomposition into

discrete parts. This is obviously, almost tautologically, true of compound symbols; and it's ***vacuously true***[103] of non-compound symbols, since no such symbol has any internal structure at all.

5.1 The picture theory of meaning (continued)

It's not entirely clear why Wittgenstein said that sentences are pictures. But it is clear that this contention of his is consistent with his empiricism.

According to empiricism, all knowledge is known strictly on the basis of sense-perception. If know it, either

(a) You directly observed it

or

(b) Without using any inferences rules of whose legitimacy your senses didn't apprise you, you inferred it from what you *did* observe,

Our sense-perceptions give us pictures. Not all sensory modalities give us *visual* pictures, of course. Hearing gives us auditory pictures, touch gives us tactile pictures, etc. But perceptual representation is pictorial representation. (In what follows, when I say "see," I mean "see or hear or touch [etc.]."

But much of what we know can't be embodied in images of any kind. (In this context, I'll use the word "image" to refer to represent moving pictures as well as still-images.) I know that:

(1) The moon is not made of cheese.

What would an image of this fact be? An image of a cheesy moon with a big X on it? No. The big X wouldn't be an *image* at all. Like the word "not," it would be a conventional sign of negation. Whereas a picture of a cheesy moon would indeed resemble a cheesy moon, a big X doesn't resemble the operation of negation. Nothing could physically resemble that operation, since it isn't something that could possibly be seen or otherwise sense-perceived.

Also, a picture of a cheesy moon corresponds to a lot of different propositions.[104] Any such picture will also depict an object having a certain color, shape, etc. Since (1) doesn't say anything about the moon's shape or color, it isn't identical with such an image. No proposition is identical with *any* image, since any image will contain information not contained in the

proposition.

Thus, there are at least some cases where one's knowing of some fact doesn't consist in there being an image in one's mind of that fact. How is the strict empiricist to deal with this? First of all, it's hard to see how sense-perception, which gives us nothing but one image after another, could apprise us of truths that are incapable of being expressed in a strictly imagistic form. For argument's sake, let's concede to the empiricist that it's strictly through perception that I know that the Moon is made of *XYZ*. How are we to deal with my subsequent knowledge that the Moon is not made of cheese? Obviously that knowledge is *largely* based on my knowledge that it's made of *XYZ*. But it can't be *entirely* based on it. What doesn't exist can't be seen. It may be that *on the basis* of what I see, I know that the moon isn't made of cheese. But I don't *see* its non-cheesiness. So *some* kind of non-perceptual knowledge is involved in my making the leap from *the moon is made of XYZ* to *the moon is not made of cheese*.

Here is a similar argument. Even if image-resistant facts (e.g., those expressed by negative statements) are learned in a strictly perceptual manner, the mental states that mediate our knowledge of them are not themselves images. The information borne by those mental states must be encoded in some non-iconic form. This means that, at some point, pictorial information was converted into non-pictorial information. But if our post-perceptual mental states are to be knowledge, that conversion process must be a legitimate one. In other words, it can't, when given pictorial input *x*, yield some output *y* that is inconsistent with *x*. Moreover, we must *know* that the conversions being made are legitimate. For argument's sake, suppose that I don't *know*, given my (let us assume) strictly perception-based knowledge that the moon is made of *XYZ*, that it is *correct* to hold that the moon is not made of cheese. In that case, to the extent that my belief that it isn't made of cheese is based on my knowledge that it's made of *XYZ*, that belief isn't knowledge. I don't *know P* if I believe it on the basis of testimony given by a source of whose reliability I have no knowledge. Uncorroborated testimony is not itself enough for knowledge. For much the same reason, if I don't know the rules that permit the derivation of non-perceptual beliefs from strictly perceptual ones, then, even if my post-perceptual beliefs are correct, they aren't knowledge.

But there couldn't possibly be any strictly perceptual way of knowing that those conversions were accurate. Those conversions, by supposition, turn

pictures into non-pictures. So our knowledge of their existence, or (*a fortiori*) of their legitimacy, cannot itself be strictly pictorial. This is the real problem with empiricism. The rules that we use to make inferences *from* perceptual experience cannot *themselves* be learned strictly on the basis of sense-perception. This will be discussed at length in Chapters 6 and 9.

5.2 The picture theory of meaning (continued)

Interestingly, Wittgenstein made points at least vaguely like these in the TLP. I quote[105]:

> In order to be able to represent logical form, we should have to be able to station ourselves outside of logic, that is to say outside the world. Sentences cannot represent logical form: it is mirrored in them. What finds its reflection in language, language cannot represent. What expresses *itself* in language, *we* cannot express by means of language. Sentences *show* the logical form of reality. They display it.
>
> Thus, supposing that "F(x)" shows that *x* occurs in its sense, "F(x)" and "G(x)" show that the same object is mentioned in both of them. Given any two sentences, if they contradict each another, their structure shows it; the same is true if one of them follows from the other. And so on. What *can* be shown, *cannot* be said.

Wittgenstein seems to be saying that we cannot correctly describe the relationship holding between our words and the facts they describe. But, if correct, that point *itself* describes that relationship, calling into question its own coherence.[106]

In any case, contrary to what Wittgenstein says, we *can* identify the logical forms of our own utterances. To identify the logical form of a statement is simply to make it clear what it means. We can do this. We do it all the time. We do it whenever we put the meanings of words into words. There are *some* qualifications to this, as we'll see in a moment, but none that redound to the credit of Wittgenstein's point.

Wittgenstein's assertion that "what *can* be shown, *cannot* be said" is obviously false. I tell you that I can do fifty push-ups. (I say "I can do fifty push-ups.") You don't believe me. So I show you that I can do fifty push-ups. (I do fifty push-ups in front of you.) And, as we just saw, that principle holds no less in connection with logical forms than in connection with a person's ability to do push-ups.

Echoing what we said earlier, although it isn't clear why Wittgenstein made these claims or what he meant by them, it is clear that they are consistent with his view that sentences are pictures of reality. A *picture* cannot picture itself. If *P* is a picture of a seagull, *P* can't contain a picture of itself, for the simple reason that nothing can be a proper part of itself. Of course, *P* might be a picture of a big seagull *and also* of some other, much littler, but otherwise identical seagull. But the big-seagull part of the picture isn't identical with the little-seagull part. The big part contains *two* seagull-images; the little part only contains one.

So *if* sentences were pictures, then a given sentence S_1 couldn't be a picture of itself. But it doesn't follow that some other picture S_2 couldn't be a picture of S_1. Nor, therefore, does Wittgenstein's much stronger claim that *nothing*—no picture, no set of pictures, no sentence—could depict or otherwise represent the rules by which true statements are paired off with the facts they described.

Wittgenstein's claim that we cannot state the logical forms of sentences, which collapses into the brazenly false claim that we can never say what our words mean, has two roots. One of them is his just-discussed belief that sentences are pictures of the facts they describe. The other is his not yet discussed belief that if it's assumed that we *can* say what our words mean, we have no way of dealing with paradoxes like the following. Consider the sentence:

(i) "what I'm saying is false."

It would seem that an utterance of that sentence is true just in case it's false, and therefore that it's both true and false---which is impossible, since whatever is true is, for that very reason, not false. (In Chapter 7, we'll find that (i) is ambiguous and that neither of its disambiguations is paradoxical in any way.) Wittgenstein was keenly interested in this paradox during his pre-Tractarian years.[107] His reaction to it, it appears, was to hold that *any* attempt to articulate semantic rules would self-refer in the same paradox-engendering way as (i) and, therefore, that such rules cannot possibly be put into words.

Wittgenstein's reasoning is not good. When I say, while pointing at the person exiting the limo, "that's Mick Jagger," I'm stating a semantic rule.

There is some individual *x* such that I am saying (correctly, we may suppose) that it's a semantic rule that "Mick Jagger" refers to *x*. The semantic rule I'm expressing doesn't self-refer and isn't otherwise defective.

As we'll see in a moment, there are reasons to think that *some* semantic rules cannot be put into words, and Wittgenstein seems to have had at least a vague knowledge of some of them. But given only that *some* semantic rules can't be put into words, it obviously doesn't follow that none of them can.

5.3 "Truth" defined

Having spent pages dwelling on the shortcomings of the *Tractatus*, let's end this section on a sunnier and more constructive note.

It is often said that truth is indefinable. Many people say this without meaning anything by it. Setting such people aside, those who say this seem to mean either:

> (a) that it cannot be said what it is for a proposition to be true,

or:

> (b) that no language can state all of the semantic rules that belong to it.

(a) is false (for reasons given at the end of Chapter 3). (b) is true.

(b) isn't the absurd claim that no language can state *any* of its own semantic rules. Every time one uses an English sentence to define an English expression, one is expressing a semantic rule of English in English. But neither the English language, nor any other, can state *all* of its own semantic rules.

Here's why. Given any meaningful expressions, there is a semantic rule *r* saying what *s*'s meaning is. (This is trivially true. A meaningful expression is one that has a meaning and is therefore one such that some true proposition identifies that meaning; and any proposition that, like *r*, says what an expression means is *ipso facto* a semantic rule.) So, for example, supposing that *x* is Dick Cheney, there is a semantic rule of English to the effect that "Dick Cheney" refers to *x*. That rule is not *itself* a sentence, but it can be expressed by a sentence. In general, semantic rules, though often capable of

being expressed by sentences, are not themselves sentences. In light of this fact, suppose for argument's sake that for each semantic rule of English there is a sentence of English that expresses that rule. Let *K* be the least inclusive class containing every sentence of English that correctly expresses a semantic rule of English. Let *SRE* be the conjunction of all of *K*'s members. *SRE* is a true and therefore meaningful sentence of English. *SRE* is also a member of *K*. After all, *K* contains every sentence that correctly says what is meant by at least one expression of English, and *SRE* obviously satisfies that requirement. But given that *SRE* is also a conjunction consisting of all of *K*'s members, it follows that *SRE* is one of its own conjuncts. No conjunction can be one of its conjuncts. (The conjunction "snow is white and snow is white" is not one of its own conjuncts, since that sentence is a conjunction, whereas "snow is white" is not.) We're forced to reject this obvious truth *if* we grant the supposition that the English language can express each of its own semantic rules. Therefore, the English language cannot state all of its own semantic rules. Given any language *L*, what we just said about English is true of *L*. Thus, no language can express all of its own semantic rules.[108]

6.0 (ii) revisited: formal truth ≠ analytic truth

One of the main contentions of the TLP is that all entailment is ***formal entailment***. One statement *entails* another if, supposing the first is true, the second couldn't possibly be false. So "Smith is a triangle" entails "Smith has more than one side." One statement, S_1, formally entails another sentence, S_2, if the statement "if S_1, then S_2" is a ***formal truth.*** A statement is a formal truth if every statement having the same form as it is true. Thus:

> (1) "If Smith is in the barn, then it is not the case that it is not the case Smith is in the barn,"

is a formal truth, since every statement of the form:

> (2) "If *P*, then it is not the case that it is not the case that *P*,"

is true.

(1) is also an example of a formal entailment. It's a formal entailment since it's a formal truth that is also an entailment. (It's an entailment since it's

to the effect that that one statement (*Smith is in the barn*) entails another (*it is not the case that it is not the case that Smith is in the barn*).

In the TLP, Wittgenstein asserted that all entailments are formal entailments. He was aware that there are apparent counterexamples to this. For:

(3) “Brown is a bachelor,”

entails that:

(4) “Brown is unmarried.”

But:

(5) “if Brown is a bachelor, then Brown is unmarried,”

isn’t formally true, since it has the same form as:

(6) “if Brown is a bachelor, then Brown is a cupcake,”

which isn’t true at all, let alone formally so.

Wittgenstein deals with this by saying, very reasonably, that (4) is synonymous with:

(3_F) “Brown is unmarried, and Brown is an adult and is male.”

Unlike (3), (3_F) *does* formally entail (4). According to Wittgenstein, all apparent counterexamples to his thesis that all entailment is formal entailment can be dealt with similarly.

But the method used in connection with this particular counterexample fails in connection with others. The sentence:

(7) “Brown is a circle,”

entails:

(8) “Brown is a two-dimensional figure.”

Since (7) obviously doesn’t formally entail (8), Wittgenstein must say that (7) is synonymous with some other expression that does. If there is *any* sentence that is synonymous with (7) that formally entails (8), it’s:

(9) “Brown is a closed, planar, two-dimensional figure of uniform curvature.”

(9) does indeed formally entail (8). But (9) isn’t synonymous with (7). That’s why:

(10) “Brown is a circle iff Brown is a circle,”

is trivial, and says nothing, whereas:

(11) “Brown is a circle iff Brown is a closed, planar, two-dimensional figure of uniform curvature,”

is non-trivial.

(11) doesn’t say anything about the *spatiotemporal* world. (11) is logically true; its truth is guaranteed by the structures of the concepts composing it. But (11) isn’t a tautology. Therefore, it’s a non-tautologous, non-empirical truth. Thus, there are non-formal entailments.

7.0 Why the concept of a logically perfect language is an incoherent one

Let’s say that a sentence is ***perspicuous*** iff its logical and grammatical forms coincide; and let’s say that a language is ***logically perfect*** iff every sentence belonging to it is perspicuous.

Many non-perspicuous sentences belong to any given natural language (e.g., English, Swedish). One of Wittgenstein’s objectives in the TLP is to identify the conditions that a language must meet if it is to be logically

perfect. Wittgenstein takes it for granted that the concept of such a language is a coherent one. Influenced by the TLP, many early analytic philosophers longed for the day when logically perfect languages would replace natural languages.[109]

In any case, the concept of a logically perfect language is not a coherent one. It isn't possible for *everything* about a sentence's meaning to be reflected in its grammar. And if *per impossibile* there did exist a logically perfect language, it would be expressively inferior to English, Arabic, and every other natural language. In other words, there would be much that couldn't be said in it that could be said in any natural language.

More precisely, for each analytic truth capable of being expressed by a logically perfect language, there would be infinitely many more that it could not express *and* that English or Spanish or any other natural language *could* express. At the same time, there would be no truth that could be expressed in a logically perfectly language that couldn't also be expressed in a logically imperfect language, such as English. Let us now discuss why this is so.

A sentence belonging to a logically perfect language is analytically true iff it is formally true. *Why is this?* A sentence is perspicuous only to the extent that its grammatical form makes it clear what it says and, therefore, makes it clear what it entails and what entails it. Bearing this in mind, let S be some arbitrary sentence. To the extent that there are false sentences, or true but non-analytic sentences, that have the same surface-structure as S, S's grammatical and logical forms don't coalesce and, consequently, S isn't perspicuous. So an analytically true sentence is perspicuous only if all other sentences having the same form are true. A sentence is formally true if, and only if, any sentence having the same form is true. Thus, a sentence is perspicuous only if formally true. Therefore, a sentence belonging to a logically perfect language is analytic only if it is formally true.[110]

A consequence is that, for each analytic truth that a logically perfect language can express, there are infinitely many that it cannot express. Given any object *x* and any property phi, the sentence

> (S_1) "‹if *x* has phi, then it is not the case that *x* does not have phi›,"

is perspicuous, the reason being that nothing having the same surface structure is false. But the superficially similar sentence:

(S_2) "‹given any object *x* and any property phi, if *x* has phi, then *x* does not have phi›,"

is *not* perspicuous, since it has the same surface structure as:

(S_3) "‹given no object *x* and no property phi, if *x* has phi, then *x* does not have phi›,"

which is false. Formally true sentences are instances of informally true universal generalizations. So no sentence capable of expressing such generalizations is logically perfect.

This point has important and often overlooked consequences. Making a ***valid*** deductive inference involves recognizing an entailment. Your deductively inferring *Q* from *P* involves your recognizing that *P* entails *Q*. The only way to know that some formal entailment is valid is to know that some informal entailment is valid. You know that:

(S_4) "Jerry is in Richmond,"

entails that:

(S_5) "it's not the case Jerry is not in Richmond."

And that's why, if you accept S_4, you also accept S_5.

But how do you know that S_4 entails S_5? Is it on the basis of your knowledge that all instances of S_1 are true? No. How could you possibly know that all of S_1's instances were true unless you could recognize the validity of specific inferences that it licenses (such as the inference from S_4 to S_5)? If you couldn't recognize that S_4 entailed S_5, then you obviously wouldn't have any idea why all of S_1's instances were correct.

This point is a psychological corollary of an epistemological point, which, in its turn, **is a corollary of a logical point. The reason why every instance of S_1 is true—in other words,** the reasonwhy S_2 is true—is that each of infinitely many specific inferences, of which the inference from S_4 entails S_5

is but a single instance, is valid. There are, quite literally, infinitely many informally valid inferences for each formally valid one. So Wittgenstein's allegation that all entailments are formal entailment is not feasible.

7.1 Hempel on the limits of strict empiricism

Carl Hempel (1905–1997) provided the following rigorous proof of the falsity of strict empiricism (what follows is not a quotation_[111]:

> (HA)[112] If x's[H1] length is one unit, and y's length is Ö2 units, then there is no length *L* such that L goes an integral number of times into both x's length and y's length. Measurement is comparison with respect to some standard. To compare x's length with y's—in other words, to establish their comparative lengths—it is necessary to find some third body z that is taken as a standard[113]; and the relative lengths of x and y are determined by finding out how many z-length segments each of x and y can be divided into. (If one object can be divided into exactly twice as many z-length segments as some other, then the first is twice as long as the second.) Supposing that x's length is one unit and y's length is Ö2 units, it follows that there is *no* body z such that both x and y can be divided, without remainder, into z-length segments.[114] It follows that, if an object's length is given by an irrational number, that fact cannot be known directly on the basis of measurement. It therefore follows that there is no strictly observation-based way to know that y's length equals x's multiplied by Ö2. In general, for any two objects x and y, there is no strictly observation-based way of establishing that x's length (or mass, etc.) is *incommensurable* with y's. (Two magnitudes M_1 and M_2 are incommensurable if there is no magnitude M_3 that goes an integral number of times into both M_1 and M_2.)
>
> The branch of mathematics known as "calculus" is integral to modern physics. Calculus is the study of continuously changing quantities. In order to describe physical phenomena in a way that makes it possible to use the powerful techniques

of the calculus to describe them, it must be assumed that they change continuously and, therefore, that the degree to which a given phenomenon has a given property may sometimes be given by an irrational number.[115] Thus, it must be assumed that, at certain junctures, the velocities, lengths, masses, etc., of objects are sometimes given by irrational numbers. But there cannot, as we've seen, be strictly observational grounds for believing that a given object's length is Ö2 meters (or that its mass is Ö2 lbs, etc.).[116] Since the calculus can't be applied to observable phenomena unless it's assumed that things' weights, velocities, etc., can at least sometimes assume values given by irrational numbers, it follows that modern physics integrally depends on an assumption for which there cannot possibly be a strictly observational basis. This means that strict empiricism is inconsistent with the obvious fact that modern physics is a source of knowledge.

Chapter 2 Properties and Non-Spatiotemporal Existence

1.0 Why we must grant that there are non-spatiotemporal entities

Not everything that exists is in space-time. Everything that is in space-time is an ***instance*** of at least one ***property,*** and properties are not themselves in space-time. I like to play tennis. I thus have the property of liking to play tennis and am thus an instance of that property. I am obviously in space-time, but that property is not. No matter how thoroughly one searches the universe, one will never encounter the property of liking to play tennis. One will, at most, encounter its instances.

Many hold that, although property-instances exist, properties *per se* either don't exist or exist but only because they're identical with spatiotemporal entities of some kind (e.g., aggregates of their own instances, ideas in people's minds). In this chapter, we'll discuss why this seemingly reasonable viewpoint cannot be sustained. We'll also see why the negation of this viewpoint is, contrary to initial appearances, utterly consonant with commonsense.

Properties themselves have properties. The property of being a square circle has the property of being uninstantiated. Properties of properties are ***higher-order properties.*** Properties that are not higher-order properties are ***first-order*** properties. The instances of higher-properties, being properties themselves, are not in space-time. The instances of first-order properties are in space-time.

Everything that exists is either a property or is in space-time. Non-spatiotemporal entities that *seem* not to be properties always turn out to be so. For example, ***propositions*** are not spatiotemporal and propositions do not initially seem to be identical with properties. (Propositions are the things that we affirm: so, setting aside cases of insincerity and linguistic incompetence, what one affirms in saying "snow is white" is a proposition, and in saying "schnee ist weiss", one affirms that same proposition.) But propositions turn out to be properties. (We'll see why in Chapter 4.) ***Sets*** are not spatiotemporal, and sets do not initially appear to be identical with properties, but they too turn out to be properties.[117] (However, contrary to what many a philosopher has said, properties are *not* identical with the sets that contain their own instances. The property of being a bird is not identical with the set of all birds. It's identical with a very different set, as we'll see at the end of the present chapter.)

The doctrine that spatiotemporal things are instances of non-

spatiotemporal things is known as “Platonism.” This is because Plato was the first person to advocate that doctrine. Sometimes Platonism is known as the *theory of universals*, and we will sometimes refer to it that way. The doctrine that properties don’t exist, or do exist but only because they are identical with spatiotemporal entities, is sometimes referred to as “nominalism.” But we’ll usually refer to it as “anti-Platonism.”

Incidentally, nominalism is so called because some anti-Platonists held that universals are identical with word. (“Nomen” is the Latin word for “name” or “word.”) In Section 3.0, we’ll see that this position, in addition to being false, is incoherent, the reason being that words are themselves universals.[118]

1.1 Why we must grant that there are non-spatiotemporalentities (continued)

Given that:

(1) Sally and Fred are both intelligent,

it follows that there is something that they have in common; that is, that:

(2) There exists some property *P* such that Fred is an instance of *P* and Sally is an instance of *P*.

Instances of intelligence are in space-time. Einstein was in space-time; so was Keynes. But intelligence itself is not in space-time. “But mightn’t the property of intelligence be nothing other than some sort of aggregate of all its instances?,” one might ask. “Mightn’t there be some way of spatiotemporalizing that entity?”

No. Consider the property of having a perfectly triangular shape. Nothing has this property. Are we to say that it doesn’t exist? Were we to do so, we’d be saying that, since there are no instances of the property of triangularity, that property doesn’t exist. Which is the same as saying there exists a property, viz., that of triangularity, such that, because nothing has it, it doesn’t exist. Which is absurd.

Also, a consequence of identifying properties with “aggregates” of their instances is that properties that clearly aren’t identical with each other must be seen as being identical with each other. (In this context, the word “aggregate” can refer to any spatiotemporal entity that is composed, in *any* sense of the word “composed,” of other spatiotemporal entities.) There are no perfectly spherical objects; nor, as we just pointed out, are there any perfectly triangular objects. So, if we identify properties with aggregates of their instances, the aggregate of all the instances of the one property is identical with the aggregate of all the instances of the other, the reason being that each aggregate is a perfect nullity.

1.2 Logic as property-based, not set-based

Consider the clearly correct statement that:

(1) Squareness is a form of rectangularity.

On the face of it, (1) seems to make a statement about the property of squareness. So, if (1)'s appearance is to be trusted, then there exists a property (namely, squareness) that has another property (namely, rectangularity).Many philosophers[119] have tried to show that, despite its appearance, (1) doesn't really talk about properties *per se* and that it instead talks about their instances. They've proposed that the real meaning of (1) is:

(2) Everything that is square is a rectangle.

And they've proposed that other statements, such as (1), that seem to refer to properties, be treated similarly. (So "green is a color," they say really just means that everything green has a color.)

But (1) doesn't mean the same thing as (2). If there were exactly two square things, *x* and *y*, and *x* and *y* happened to be green, the statement:

(3) Everything that is square is green,

would be true, but:

(4) Squareness is a form of greenness,

would be false (and so would other statements of its ilk; e.g., "squareness is a color").

Of course, the statement:

(5) Everything that is square is *for that very reason* a rectangle,

is true, whereas:

(6) Everything that is square is *for that very reason* green,

is false, even if the only two square things in existence happen to be green.

What (5) says is that:

(7) A thing's being a square is responsible for its being a rectangle,

which is obviously correct; and (6) says that:

(8) A thing's being a square is responsible for its being green,

which is obviously false.

But (5) collapses into (1). (5) says that, to the extent that everything square is rectangular, it's because the property of squareness bears a certain relation (that of being a form of) to the property of rectangularity. The relationship between a thing's being a square and its being a rectangle is, (5) rightly says, to be understood in terms of a relationship holding between squareness and rectangularity. But the relationship between a thing's being square and its being *green* is not to be understood in such terms.

A thing is a rectangle *in virtue of* its being a square. But, given some object that is both green and square-shaped, it isn't *in virtue* of its being green that it's square-shaped. Whenever one encounters an expression like "in virtue of," "*ipso facto*," "to the extent that," or "for that very reason," one knows that one is dealing with a statement that is primarily about properties and only secondarily about objects. If I say that "copper things are *ipso facto* good conductors of electricity," I'm saying that the relationship between a given thing's being made of copper and its being a good conductor of electricity isn't fortuitous and is grounded in the natures of those two properties. In other words, to understand why an instance of the first property is an instance of the second, one looks to the natures of those properties. The connection goes from property instance (instance of being made of copper) to property (property of being made of copper) to property (property of being a good conductor of electricity) to property-instance (instance of being a good conductor of electricity).

And the reason why "copper things are *ipso facto* in JMK's living room" is false is that, even though (so we will suppose for argument's sake) every copper object happens to be in my living room, the connection between *x*'s being made of copper and *x*'s being in my living room isn't routed in that way through those properties.

All *principled* connections are relationships between properties. This means that ***nomic***[120] connections (e.g., the connection between a thing's being made of copper and its being a good conductor of electricity) and also all logical connections (e.g., the connections between a thing's having knowledge and its having a mind) are expressed by statements whose subjects are properties. The correct statement that:

> (9) For any *x*, the statement *x has knowledge* entails *x has a mind*,

is another way of saying that:

> (10) Something that has knowledge *for that very reason* has a mind of some kind.

And (10) is another way of saying that:

> (11) The property of having knowledge is a form of the property of having a mind.

In conclusion, unless it's conceded that properties exist, there is no way to account for the fact that there are *principled* relationships of any kind. And in fact those who deny that properties exist tend *not* to grant that there exist any such relationships. Those who deny the existence of properties tend to be ***empiricists.*** Since empiricists cannot grant the existence of things, such as properties, that cannot be seen or otherwise known (more or less) directly on the basis of observation, they end up *denying* that anything has any bearing on anything. (See Chapter 18.)

1.3 Logic property-based, not set-based (continued)

Let S_1 be the set containing the smallest number greater than zero that is divisible by two, and let S_2 be Fred's favorite number, which happens to be the number two. S_1 and S_2 are the same set, and the statement:

(1) *x* is a member of S_1,

is therefore identical with the statement:

(2) *x* is a member of the set that contains the number two and doesn't contain anything else;

and both statements are identical with:

(3) *x* is a member of S_2.

Thus, anything entailed by any one of those statements is entailed by each of the others, and anything that entails any one of those statements entails each of the others. Thus, (1) entails:

(4) *x* is less than twenty,

and so do (2) and (3). And:

(5) *x* is a whole number greater than one but less than three,

entails (1), as well as each of (2) and (3).
But the statement:

(6) *x* has the property of being Fred's favorite number,

entails:

(7) *x* has the property that Fred is aware of its existence,

but (1) doesn't entail this, and neither do (2) or (3). And

> (8) *x* has the property that it is a number that is preferred by Fred to every other number that Fred has considered,

entails (6), but it doesn't entail (1) or (2) or (3). It follows that properties cannot be identified with sets of their instances without mutilating the most basic principles of logic.

"But surely that's wrong," one will say. "(1) is the statement that *x* is a member of the smallest set that contains the smallest number divisible by two," and (2) is the statement that *x* is a member of the smallest set containing Fred's favorite number. So I don't see how (1) and (2) could entail, and be entailed by, the same things; nor do I see how (1) and (6) could entail, or be entailed by, different things."

If the expression "the set containing Fred's favorite number" is taken in this way—in other words, if it's taken in such a way that (1) and (6) are equivalent, whereas (1) and (2) are not—then the word "set" is going proxy for the word "property." If by "the set containing Fred's favorite number" you mean, *not* "the set containing the number two," but instead "*any* set, whatever its membership, that comprises Fred's favorite number," then what you mean by "the set containing Fred's favorite number" is: "any set that is generated by the *property* of being Fred's favorite number." But in that case, when you make statements about "sets," you're really talking about properties.

1.4 The axiom of extensionality

It may be obvious at this point that properties aren't identical with sets of their instances. But this now obvious point is easily extended to reveal the futility of attempts to show that properties, if existent, are spatiotemporal. But in order to make the needed extensions, we must take a very quick detour through a bit of set theory.

Two very different properties can be ***co-extensive.*** Two properties are co-extensive if any instance of the one is an instance of the other. But different *sets* cannot have the same members. This is known as the ***axiom of extensionality.*** Since different properties can generate the same set, properties aren't identical with sets of their instances. We'll revisit this point later in this chapter. So if S_1 is a set and S_2 is a set, then S_1 is identical with S_2 if and only if nothing is a member of the one that isn't also a member of the other. A set whose members are *x*, *y*, and *z* can't fail to have those members any more than it can fail to be itself—indeed, for it to fail to have those members would be for it to be fail to be itself. Thus, sets are *individuated* by their memberships. This principle is sometimes known as the axiom of extensionality.

A corollary of this axiom is that a set's membership is frozen—frozen, indeed, in two senses. First, it's temporally frozen. Sets don't gain members and they don't lose them. Second, it's *modally* frozen. If a set's actual members are $x_1...x_n$, it couldn't possibly exist without having exactly those members.

This may at first seem patently false. "Surely the set of people changes from year to year; people are born; people die." This is the wrong view. The right view is that some one *property*—that of being a person—generates *different* sets at different times. So really there is no such thing as *the* set of people. There is the set of people at 3:00 P.M. April[H2] 28, 2009, the set of people at 3:01 P.M. April 28, 2009, etc.

And if *S* is the set of people—in other words, if it's the set that contains you, me, Barack Obama, etc.—*S* couldn't possibly exist in a universe in which one of us three didn't exist. Of course, there could be *people* in a world where one of us didn't exist. And, in that world, the *property* of being a person would generate *a* set. But that set wouldn't be *S*. It would be some other set.

"But," one might ask, "isn't the axiom of extensionality just a matter of convention? Couldn't we stipulate that sets *could* lose and gain members?" If we used the word "set" in this way, it would become a proxy for the word "property." To say that the "set" of humans changes every time someone is born or dies is to say that, at any given time, this set's membership is determined by what it is that, at that time, has the *property* of being human. So, while there's nothing wrong with using the word "set" in this way, it isn't an option for somebody who wishes to replace properties with sets.

Having taken care of these technical preliminaries let us move on to matters of substance. If some piece of copper happens, at a given time, to be conducting electricity, it's because of what is going on right then and there. The fact that there exists some other piece of copper in some remote galaxy is irrelevant. But if *S* is the set of all pieces of copper, the existence of that other piece of copper *is* relevant to *S*'s existence. Sets are individuated by their memberships. (The set consisting of Smith, Larry, and Sally ceases to exist if Smith does.) So if, for some reason, the universe hadn't begotten that other piece of copper, then *S* wouldn't exist. But what difference would that make to the fact that *this* piece of copper is conducting electricity? None.[121]

Any given thing belongs to many different sets. What those sets are is only as good an indicator of how that thing is as it is an indicator of what properties that thing has. If *S* is the set of people who actually exist, then *S* doesn't exist in a world *W* where Dick Cheney doesn't exist but is otherwise exactly like our world. But what difference does that make to my life-path in *W*? None. So far as you can learn anything about me from the fact that I belong to the set of humans, it's because, in learning this, you are learning that I have the property of being human.

Everything just said about sets can be said of *any* entity that is constituted by property-instances. Suppose, for argument's sake, that there is a "scattered object" consisting of all and only those human beings in existence. (This is a popular view among philosophers.[122]) Just replace each occurrence of "set" in the argument just given with "scattered object consisting of all and only human beings," and the resulting argument shows that the property of being a person isn't identical with that set. (Also, we'll see in the next section that, setting aside properties that have only one instance, there are *no* spatiotemporal entities that can be identified with a given property's instances. There are no "scattered objects.")

2.0 Arguments against Platonism

It is seldom denied that there are instances of properties, but it has often been denied that there are non-spatiotemporal entities. But it's hard to see:

> (i) How there could be instances of properties without there being properties themselves,

and it is also hard to see:

> (ii) How properties could be given spatiotemporal coordinates.

Opponents of Platonism respond either by saying that (i) involves some sort of logical or linguistic trickery or by saying that properties *can* be identified with spatiotemporal entities. Let us now consider how anti-Platonists defend their view.

2.1. Argument #1 against Platonism

> Not every word in a sentence corresponds to some entity in the world. If I say "for the sake of all that is good, make sure you give at least some money to charity" the word "sake" doesn't denote anything—there aren't such things as sakes, after all—even though, being a noun, the word "sake" has the same syntactic function as words that do have denotations.
>
> We can develop this point by taking a brief tour through linguistic theory. It seems fairly clear that not all expressions *refer* to things.
>
> Consider words like "of," "and," "the," "nonetheless." It's hard to believe that these words are comparable to "Smith," "Jones," and "cow." The medievals referred to expressions like "Smith" and "Jones" as "categorematic" expressions, and to expressions like "of" and "the" as "syncategorematic" expressions. A syncategorematic expression is one that doesn't refer to anything and whose linguistic function is strictly syntactical. Contemporary linguists make a similar distinction: they distinguish between "lexical" and "non-lexical" items: the

former include “Smith,” “Jones,” “red,” and “cow”; the latter include “the,” “of,” and “nonetheless.” A lexical item is one that has a referent; a non-lexical item is one that does not, and whose purpose is strictly syntactic. (In other words, the purpose of a non-lexical term is simply to facilitate the unification of lexical items to form larger lexical items.)

Even though there are many expressions that appear to denote properties, they are really syncategorematic or non-lexical expressions. Words like “red” or “the property of redness” don’t refer to anything. This is not to say that they are not meaningful. “Of” is meaningful, even though it doesn’t denote anything, and the same is true of “the property of redness.”[123]

***This argument evaluated*:** If I know that:

(1) *x* is red,

is true, I can logically infer:

(2) There exists some color or other that *x* has.

If I know that:

(3) Fred is smart but Larry is not,

I can infer that:

(4) There exists some property that Fred has that Larry lacks.

These inferences are obviously valid. The view that “red” and “smart” are not syncategorematic invalidates them. Therefore, that view is false. So, just as the Platonist holds, “red” and “smart” pick out entities of some sort or other. To use the objector’s own (Medieval) terminology, those expressions are “categorematic,” not “syncategorematic.”

When we make it clear just what these terms mean, it becomes clear that the objector’s argument not only fails, but fails relative to its own

assumptions. A syncategorematic entity is one whose function is to form complex expressions out of other expressions. Thus “and” is a syncategorematic entity since it enables complex expressions (“snow is white and grass is green”) to be formed out of other expressions (“snow is white”,” “grass is green”). “Of” is a syncategorematic expression for the same reason: given the expressions “picture” and “Paul”, “of” yields the complex expression “picture of Paul.”

Bearing this in mind, consider the sentence:

(4) “snow is white and grass is green.”

Although “and” serves a purely “syntactic” function, it does so, at least in part, *by way of* picking out a certain (Platonic) entity. This is because (4) entails:

(5) There is some relation R such that the proposition that snow is white bears R with respect to the proposition that grass is green (and *vice versa*).

What we said about the occurrence of “and” in (4) is true of the occurrence of “of” in:

(6) There is a picture of Paul on the mantelpiece.

This is because (6) entails that:

(7) There is some relation R such that there is a picture on the mantelpiece that bears R with respect to Paul.

So even syncategorematic items sometimes have referents. Therefore, even if “red” and other predicates were syncategorematic, it wouldn’t follow that predicates didn’t have referents.

Gottlob Frege (1944-1925) put forth the following (italicized) argument, which some authors have taken to show that predicates don’t refer to anything and, therefore, don’t refer to properties (“NP” is short for “No

Properties”).

> *(NP) The expression “the property of whiteness” refers to the property of whiteness, at least if anything so refers. So if the occurrence of “white” in:*
>
> *(8) “snow is white”,*
>
> *referred to that property, then (8) would have the meaning as:*
>
> *(9) “snow, the property of whiteness.”*
>
> *But (8) is meaningful, whereas (9) is meaningless; so they don’t have the same meaning.*

Although Frege himself was an ardent Platonist, many have taken NP to show that properties don’t exist. Here is their reasoning:

> (NP*) If there is any reason at all to posit the existence of properties, it is to explain the meaningfulness of predicates. Given that predicates don’t refer to properties, as we’ve just seen, but are still meaningful, there is no reason to suppose that properties exist.

NP *does* show that “red” doesn’t *only* refer to the property of redness; it doesn’t show that referring to that property is *no* part of what “red” does. The expressions “Bill is a member of the species *homo sapiens*” and “A member of the species *homo sapiens*, Bill is a scoundrel” cannot be inter-substituted. But the former *does* say that Bill is a human. A given occurrence of “he” can co-refer with a given occurrence of “him,” even though, supposing that each occurs in a grammatical utterance, neither occurrence can replace the other without turning a meaningful expression into a meaningless one. So NP doesn’t show that “white” doesn’t refer to the property of whiteness.

That said, NP does prove an important point: that not all meaning takes the form of reference. (9) is not a sentence. In general, a list of referring terms

isn’t a sentence. And, given a list of referring terms that isn’t a sentence, adding referring terms to that list won’t turn it into a sentence. Adding “the relation of having” to (9) yields:

> (10) “snow, the property of whiteness, “the relation of having,”

no permutation of which is a sentence.

There is thus a kind of meaning that has nothing to do with reference. The inflections in (8) have this non-referring sort of meaning, the nature of which we’ll clarify in Chapter 3.

Unlike NP, NP* is a complete failure. There are many reasons to hold that properties exist, and most of them have nothing to do with language, as we’ve already seen.

2.2 Argument #2

Although properties exist, they are identical with spatiotemporal entities. The property of being a rock is identical with the *set* of rocks. And the set of rocks is a spatiotemporal entity. After all, that setis simply a very spread out object that consists of rocks.

***This argument evaluated*:** Sets are abstract objects. They don't have space-time coordinates. The *set* of rocks is not some big rock and it isn't some big object consisting of rocks separated by stretches of empty space. If all the rocks in the world move to some new place, and none are destroyed, the set of rocks hasn't changed at all: it is the same as it was before. But, under those circumstances, any object consisting of all the rocks in the world *has* changed: it has acquired a new location (or, in any case, it has undergone an internal structural change).

A set is a mathematical entity; it has no location and is defined entirely by its formal properties. *Where* its members are and what they are doing, is irrelevant (so long as its members don't cease to exist). But any object that is actually *composed* of all the rocks in the world *is* affected by the movements of their members, and thus cannot be identified with sets. (Of course, it is unclear whether there is such an object—an issue we'll revisit shortly—but what is clear is that *if* there is, *then* it is affected by movements of the kind just described.) Thus, sets aren't spatiotemporal entities. Consequently, they are *themselves* Platonic entities. So, one is conceding the truth of Platonism, not undermining it, by identifying properties with sets.

Aware that, for the reasons just given, properties cannot be identified with sets, philosophers of an anti-Platonist bent have sometimes chosen to identify properties with what are referred to by contemporary philosophers as *scattered objects.*

First of all, what exactly is a scattered object? Ordinary objects are geographically and dynamically cohesive. Consider your back-pack. If you move one of the straps ten feet across the room, you have moved the whole bag across the room (unless, of course, you've damaged the bag in some way). And even though (at the microscopic level) the bag consists mostly of empty space, all of its parts occupy one, fairly compact, relatively continuous region of space-time.

Now consider all the rocks in the universe. Those rocks clearly *don't*

constitute a dynamically or geographically integrated unit. Supposing that those rocks collectively constitute any object at all, the object they constitute is a "scattered object." In general, a scattered object is one that lacks the cohesiveness and dynamic integrity of ordinary objects. According to some anti-Platonists, properties are scattered objects. Is this correct?

No. There are no scattered objects. The very concept of such a thing is an incoherent one. In Chapter 20, we'll see why. Here's an outline of what we'll say. It's obvious that two objects that are spatially separated from each other can both be constituents of some third object. But if spatially disjoint entities are to constitute a single spatiotemporal object, their interactions must be characterized by a certain regularity and predictability. Dynamic cohesiveness must step into the breach opened by a lack of spatial cohesiveness. Since it cannot be known without doing empirical research whether the various rocks in the universe jointly constitute an entity have the requisite cohesiveness, it cannot be said on non-empirical grounds whether there is some "scattered object" consisting of all the rocks in existence. But, clearly, the property of being a rock *does* exist. So[H3] that scattered object, even if it does exist, isn't identical with that property. And that scattered object doesn't exist: given the empirical evidence at our disposal, there is no real chance that the various rocks in the universe, taken together, have the dynamic integrity that is a *sine qua non* for being a veritable object.

But even if scattered objects do exist, properties aren't identical with them. Let *R* be some scattered object composed of all the rocks in the universe. If you destroy half of the rocks in the world, you have changed the just-mentioned scattered object. But you haven't changed the conditions that a thing must satisfy to be a rock. (For argument's sake, we'll suppose that to be a rock is to be hunk of matter with certain microstructural properties.) Therefore, you haven't changed what it is to be a rock or, consequently, the property of being a rock. So, by Leibniz's Law, that scattered object, if it even exists, cannot be identified with the property of being a rock. Given any other property, the same points *mutatis mutandis* show it not to be a scattered object.

2.3 Argument #3: The third-man argument

According to Platonism, any given spatiotemporal entity is an instance of at least one property, and for two spatiotemporal entities to resemble each other in some respect is for there to be some property of which each is an instance. Presumably, properties must resemble their instances. For example, the property of being triangular must itself be triangular. Bearing this in mind, let *T* be the property of being triangular, and let *x* be an arbitrary triangle. If Platonism is correct, it follows that there is some property T_1 such that *T* and *x* are both instances of T_1 and such that it is in virtue of their both being instances of T_1 that they are triangular. Given that, according to Platonism, properties resemble their instances, it follows that T_1 must resemble each of *T* and *x*. In other words, T_1 must be a triangle. It follows that, if Platonism is correct, there must be some property T_2 such that each of T_1 and *T* and *x* is an instance of T_2 and such that it is in virtue of their being instances of T_2 that they are triangular. And so on ad infinitum. Were they to exist, properties would be useless, since the very questions that arise in connection with their instances arise in connection with them.[124]

This argument evaluated: Consider the property of being spatiotemporal. If Platonism is right, that property is not itself spatiotemporal and, in the relevant sense of "resemble," it doesn't resemble its instances. Here is another example. To be triangular is to have a certain shape. Only things that are in space have shapes. The property of triangularity is not in space and, therefore, doesn't have a shape and, therefore, isn't an instance of itself. TMA succeeds in showing that instantiation isn't resemblance. This means that it succeeds in proving nothing, given that it's inherent in Platonism that instantiation *not* be resemblance.

2.4 Argument #4: The argument from the causal impotence of properties

Properties lack causal powers. For this reason, they can serve no explanatory function. To explain something is to show how it came about. Being causally impotent, properties can't make anything come about. Therefore, there is no reason to posit them

This argument evaluated: Properties do indeed lack causal powers. Causes are necessarily in space and time. To be a cause is to induce change. Changes and their causes are *ipso facto* in space-time.

But this is no reason to think that properties are explanatory deadwood or, therefore, that there's no good reason to believe them to exist. Not all explanation is *causal* explanation. Some explanation is what we call *analytic* or *conceptual*. There are as many even numbers as there are whole numbers. (This is a theorem of set-theory.) How is this to be explained? Even though the even numbers are a subset of the whole numbers, it is possible to put the two sets of numbers into a one-one[125] correspondence. Since, whenever two sets can be put into such a correspondence they have the same number of members, it follows that the set of even numbers has as many members as the set of whole numbers.[126]

This explanation is not *causal*. It explains a truth *not* by identifying its causal antecedents, but by delineating the structures of concepts. Many explanations are like this. You ask me: "Why can't someone know something without having any form of awareness?" I answer (correctly) by saying: "because knowing something involves having a true belief, and a belief is a form of awareness." This explanation identifies conditions that are conceptually, not causally, necessary conditions for knowledge. Like this explanation, Plato's theory isn't a causal theory, and it's therefore irrelevant that properties aren't in space-time.

2.5 Argument #5 (a variant of #4): The argument from the causal theory of knowledge

For *x* to know of *y*, *y* must have causal *effects* on *x*. You can see the dog only because the dog has certain effects on you—light bounces off its fur and hits your eyes. You can know about Socrates only because, long ago, Socrates made various noises, which were transcribed, and copies of those transcripts made their way to you. So it is only because there is a causal process linking Socrates to you that you know about him and are able to have thoughts about him.

Properties don't have effects on anything. They don't initiate or otherwise participate in causal processes. So we can't know about them, even if they exist. So there is no good reason to suppose that they exist.[127]

This argument evaluated: Seeing a dog does indeed involve the dog having effects on you—it involves your eyes being disturbed by photons bouncing off of the dog. But, in and of themselves, those photonic disturbances are lifeless and informationally empty. If they were to hit a rock, vision wouldn't result. The photons responsible for those disturbances don't contain little dog images. You have to *read such information into them.* Your mind must extract such images from them; and to do this, your mind must have some way of interpreting those disturbances. The knowledge needed to interpret these disturbances cannot come from sense-perception, since that knowledge is itself a prerequisite for sense-perception. This entails that, in at least some cases, it isn't through a causal process that the knower becomes aware of the known. It also entails that a prerequisite for the very existence of sense-perceptions, or indeed of any mental activity that results from causal exchanges with the world, is knowledge of properties. Why does it entail this? Because, were it not for a grasp of what it is for a thing to have the property of having this or that shape (or this or that color or this or that location, etc.), one couldn't interpret the radiation-induced disturbances of one's sensory receptors in the way needed to generate sense-perceptions.

2.6. Argument #6: Conceptualism

If they exist, properties are identical with mental representations of some kind or other—presumably, with thoughts *about* rocks or, in any case, with categories created by the mind that help it upload and process sensory information.[128]

This argument evaluated: Since conceptualism is thoroughly evaluated in Chapter 13, I'll be brief. There were rocks before there people. And, back then, rocks were similar to one another in ways that they weren't similar to bats and stars and other non-rocks. For two things to be similar to one another in a certain respect is for there to be some property that they both have. So for two rocks to resemble each other in some respect is for there to exist some property that they both have. Thus, before there were human minds, there were properties. Therefore, properties are not products of human artifice.

2.7 Argument #7: Resemblance nominalism (RN)

Let *x* be some red thing. *x*'s being red consists, not in its being related in some way or other to some non-spatiotemporal entity but in its resembling roses, spots of blood, stop-signs, and other such things. Being red is a relation, not between the spatiotemporal and the non-spatiotemporal but between the spatiotemporal and the spatiotemporal. In general, to have a property phi is to bear some relation to some other spatiotemporal entity; it isn't to bear some relation to some non-spatiotemporal entity.[129]

This argument evaluated: For two otherwise dissimilar objects to have the same shape is for there to exist some shape, and therefore some property, that they both have. For two otherwise dissimilar red objects to be the same color is for there to exist some color, and therefore some property, that they both have. There is no such thing as *mere* resemblance; there is only resemblance in this or that respect. For two things to resemble each other in a given respect is for there to exist a property that they share, and resemblance nominalism therefore collapses.

2.8 Argument #8: Conventionalism

The only non-empirical truths are linguistic conventions (e.g., “there are three feet in a yard”). Therefore, all non-conventional truths are empirical. Empirical truths register spatiotemporal facts. Conventions, whether linguistic or not, are obviously spatiotemporal entities. Therefore, all truths are spatiotemporal.[130]

This argument evaluated: Since conventionalism is evaluated in Chapters 1 and 9, we can be brief.[131] Not all non-empirical truths are conventions. It’s a non-empirical truth that there are continuous functions that cannot be differentiated at any point. But it’s not a convention. *Given* the symbolic conventions that govern our usage of the relevant words (“continuous,” “function,” “differentiated,” etc.), it *follows* that the sentence “there are continuous functions that cannot be differentiated at any point” expresses a truth. But this means that it holds in virtue of convention *plus* some various non-conventional facts. Those facts are not empirical; they don’t concern the spatiotemporal world. Therefore, the fact described by that sentence is not spatiotemporal. Further, since that fact concerns logical structures of properties (e.g., the property of being differentiable), it follows that there are properties. Since, as we just noted, that fact is non-spatiotemporal, it follows that properties are non-spatiotemporal.

3.0 Nominalism

Many medieval philosophers, and some 20th century philosophers, held that properties are identical with words and, therefore, that properties are not spatiotemporal.

But this position is incoherent, since words and linguistic expressions generally are *themselves* universals or properties. How many words are right below this sentence?

> “snow”
> “snow”
> “snow”

Is it one or three? The answer: there are three *instances* of one *word*. To put it another way, there are three *tokens* of one word-*type*.[132] *Instances* of an expression are ***expression-tokens*** and the expressions themselves are ***expression-types***.

No one utterance of the word “snow” is identical with that word. It’s not as though the word-type “snow” will disappear when the inscriptions of it in the lines above this are destroyed. That word-type is a *property* of certain deposits of ink (e.g., the ones above) and bursts of noise, and those ink-deposits and bursts of noise are therefore instances of that property. This shows that, even if the nominalist is right to identify the property of being a rock with the word “rock,” he has still conceded the truth of Platonism.

Plus, that property isn’t identical with that word. The Spanish word for rock is “piedra.” It would be arbitrary to identify that property with the word “rock” but not with the word “piedra.” Since “rock” and “piedra” are obviously different words, the nominalist must say that the property of being a rock is not identical with itself, which is absurd.

All of these points aside, it makes no sense to identify that property, or any other, with a word. A thing’s being a rock has nothing to do with words; there were rocks before there were words.

4.0 Properties demystified

David Armstrong (1989) says that properties are "ways that things can be." To be round is to be one way; to be square is another. Armstrong's conception of what properties are is a fruitful one; and if it is kept in mind that properties are just ways that things can be, the resistances that we have to accepting their existence crumble.

Nobody denies that there are ways things can be. To say that there are ways things can be is to say that there are possibilities as to how things can be. Not all of those possibilities are actualized. Not everything that *could be* the case is the case. Now, actualized possibilities are in space-time.[133] But *unactualized* possibilities— possibilities that never "came true"—are surely *not* in space-time. So where are they?

One could say that they don't exist. But that would be absurd. It's one thing to say that a possibility wasn't (or won't be) *actualized;* it's a very different thing to say that the possibility *itself* doesn't exist. It's possible that you will you go to law school. But let's suppose that, even though, given our stellar grades and LSAT scores, there is clearly a sense in which law school is a possibility for you, you end up *not* going to law school. Does the fact that you *won't* go to law school mean that it isn't an option? Surely not. Does the fact that this possibility isn't actualized mean that the possibility itself doesn't exist? Surely not.

Your situation is very different from that of somebody (e.g., somebody on death-row) for whom it simply isn't an option. The fact that a path isn't taken doesn't mean that the path doesn't *exist;* and the fact that a possibility isn't actualized doesn't mean that the possibility *itself* doesn't exist. So, given that possibilities, even unactualized ones, *do* exist, we must again ask: where are they? They're not in space-time. The *possibility* of your going to law school is not in space-time. So that possibility, being outside of space-time and thus being without location, is nowhere.

At first this may seem alarming, but it ceases to be so if we choose to see possibilities as properties and actualizations of possibilities as property-instances. It thus seems that the presumptive existence of unactualized possibilities provides support for the theory of universals.

A story may help clarify the nature of properties. You are a meat inspector. You've just inspected all the meat in slaughterhouse *X*. None of it passes inspection. All of it was completely rotten. There is obviously a way

something *would* have to be to be a piece of meat that passed inspection; and there is obviously a way something would have to be to be a piece of meat *in slaughterhouse X* that passed inspection. So when reporting to your supervisor, you say "none of the pieces of meat in *X* deserved to pass inspection," you're saying: "none of those pieces of meat was the right way." Which is the same as saying: "there exists a way that something would have to be in order to be a piece of meat in *X* that deserved to pass inspection; but nothing is that way." Which is the same as saying: "the property of being a piece of meat in *X* that deserved to pass inspection is uninstantiated."

Thus, the property of being a piece of meat in *X* that deserves to pass inspection isn't a giant piece of meat in the sky. It's simply a way something would have to be in order to be a piece of meat in *X* that passed inspection.

The same is true of the property of being a pink elephant. There are no instances of that property, since there are no pink elephants. But there is a way that something *would* have to be to be a pink elephant. I know what that way is, and so do you. That's why we both know that, by going to the zoo and spray-painting one of the elephants there, we could *make* something be that way; and that's why we also know that nothing is that way (or, in any case, that nothing we've ever heard of is that way—maybe some prankster has spray painted an elephant). For something to "instantiate the property of being a pink elephant" is simply for it to be that way. So that property isn't some pink elephant in the sky. It's just a way something would have to be to be a pink elephant. And for there to be no instances of that property is simply for nothing to be that way.

Properties can be thought of as conditions. The property of being a pink elephant is a condition a thing must fulfill in *order* to be a pink elephant. Lots of unfulfilled conditions exist. Mary won't marry someone unless he fulfills the following condition: he's a genius who is always friendly and never sleeps and always dotes on her and has a billion dollars. That condition isn't fulfilled by anyone. But the condition exists. Thus, properties that aren't instantiated are just unfulfilled conditions.

So given that there are unfulfilled conditions, it follows that there are uninstantiated properties. And given that properties exist, even when they have no instances, it follows that, when they do exist, they aren't identical with their instances. The property of being a pink elephant exists, but is currently uninstantiated. If there were to come into being a single pink elephant, which promptly died, the property wouldn't go out of existence. It

would still be around. But if it were identical with instances, then it would go out of existence the moment its one instance did.

5.0 Relations and higher-order properties

We pointed out earlier that properties have properties. The property of being a bird has the property that there are more instances of it than the property of being a penguin. Actually, any statement can be interpreted as a statement about higher-order properties. “Smith has a dog” is equivalent to “the property of having a dog has the property of being instantiated by Smith.”

Higher-order properties are indispensable to scientific and ordinary discourse. If you say “true integrity is rare,” you’re saying that the property of having true integrity has the property of being seldom instantiated. If you say that ambition is more important to success than ability, you’re saying that the property of being ambitious has the property of comparing favorably with the property of having ability in respect to whether, in virtue of having it, its instances are likely to succeed.

Gottlob Frege (1844–1925) discovered that statements about numbers are statements about higher-order properties. “There are zero penguins in the house” means: the property of being a penguin in the house has the property of being uninstantiated. “There are two penguins in the house” says that the property of being a pair of penguins in the house is instantiated. The number two may be defined as what all pairs of objects have in common. For x to be a pair of objects is for it to be a set S such that there are objects O_1 and O_2, distinct from each other, that belong to S and such that anything O_2 that belongs to it is identical to either O_1 or O_2. The number two is the property of being such a pair. See Chapter 7.

Relations may be identified with properties of ordered pairs. There is some property P such that x is taller than y exactly if the ordered pair $\langle x, y \rangle$ has P. This conception of what relations are may seem contrived. But it’s an extremely useful one, since, without it, the logical forms of many inferences couldn’t be identified.

6.0 Sets and propositions identical with properties

We've seen that properties are not identical with sets of their instances. The main reason for this, we saw, is that properties tend to be modally elastic, whereas sets are modally frozen. In other words, a given property (e.g., the property of being a poet) can generate different instances in different worlds, whereas the *set* of poets is frozen. If Smith is a member of the set of poets in this world, and *S* is the set of poets in this world, then *S* doesn't exist in any world where Smith does not (or where Smith does exist but isn't a poet).

But sets still may be identified with properties. Suppose that Smith, Jones, and Brown are the only three poets in existence. Let *S* be the set of actual poets—that is, the set containing Smith, Jones, Brown, and nothing else. In that case, the statement:

(1) *x* is a member of *S*,

is equivalent with:

(2) *x* is identical with Smith or Jones or Brown,

and therefore, with:

(3) *x* has the property of being identical with Smith or Jones or Brown.

Thus, S may be identified with the property of being identical with Smith or Jones or Brown; and membership in S may be identified with possession of that property.

Sets are non-spatiotemporal entities. Given that sets are really properties, it follows that, given only that sets are non-spatiotemporal, it doesn't follow that anything non-spatiotemporal isn't a property.

Propositions are non-spatiotemporal entities. But they too are properties. The proposition *JM is playing tennis at time t* is true just in case certain properties are instantiated at certain times, and we may therefore identify that proposition with the set containing the relevant properties and identify that proposition's being true with the fact of all of that set's members' being instantiated. (In Chapter 4, we clarify and defend this point.) Since sets can,

in the way just described, be identified with properties, it follows that propositions are properties.

Properties aren't mind-independent and neither (therefore) are propositions. But, of course, it's a popular belief, mainly among empiricists, that propositions, if existent, are human creations. Here is a reason, distinct from the one just given, to think otherwise.

There are infinitely many propositions. (Since there are infinitely many locations, there are, for any given particle *x*, infinite propositions of the form *x is in place y at time t.*) We couldn't therefore have directly created every proposition there is. So, supposing that propositions are human creations, we must be only *indirectly* responsible for infinitely many of them, and it must therefore be supposed that we created some finite stock of propositions that somehow brought about the rest. But in that case there are true propositions, which we'll refer to as "intermediary" propositions, that describe how the ones we directly created brought about the others. But since there are infinitely many propositions that we didn't create, there are infinitely many intermediary propositions: one for each proposition not directly created.

Suppose for argument's sake that we didn't directly create the proposition that dinosaurs wore tweed, and also suppose that $p_1...p_n$ are the propositions that we directly created. In that case, there is a true intermediary proposition to the effect that $p_1...p_n$ brought about the propositions that dinosaurs wear tweed, another intermediary proposition to the effect that $p_1...p_n$ brought about the proposition that *either* dinosaurs wore tweed *or* 1 + 1 = 2, and another to the effect that $p_1...p_n$ brought about the proposition that it's possible that it's the case that *either* dinosaurs wore tweed *or* 1 + 1 = 2, and so on.

We didn't create each of these individually many intermediary propositions. (Given that we didn't directly create the propositions that dinosaurs wore tweed, we also didn't directly create any proposition having it as a proper part—e.g., the proposition that *either* dinosaurs wore tweed *or* 1 + 1 = 2.) We must therefore posit a second group of intermediary propositions that will mediate between the first group and the propositions that we directly created.

But everything just said about the first group of intermediary propositions will hold of the second, and we'll be no further along. Thus, it cannot coherently be supposed that we directly created every proposition in

existence. This might appear to leave open the possibility that some propositions, but not others, are human artifacts. But such a view would be arbitrary and implausible, like the view that some whole numbers are human creations while others are not. It would also be demonstrably false. We've seen that we don't directly or indirectly create the propositions describing how those that we do directly create, supposing that there are any, give rise to those we don't. Those that we don't directly create demonstrably entail those that we do.

Suppose that we didn't create the proposition that dinosaurs wore tweed but that we did create the proposition that snow is white. In that case, we didn't create the proposition that *dinosaurs wore tweed and snow is white.* That proposition entails that snow is white and thus requires its existence. Therefore the proposition that snow is white doesn't need us to create it, since its existence follows from one that we didn't create. Therefore, we *couldn't* have created the proposition that snow is white, since one can't create what already exists.

7.0 Concepts vs. properties

Some authors distinguish properties from what they refer to as "concepts." The sentence:

(BS_1) "Bill is a square circle,"

is obviously meaningful. (It's false; therefore it's meaningful.) But does the expression "square circle" pick out a property? The right answer would seem to be "no." Properties are "ways things can be." (Or so I've argued.) The laws of logic prohibit anything from being both a square and a circle. So the term "square circle" doesn't pick out a way anything can be and therefore doesn't pick out a property.

But it's meaningful. So we have a meaningful predicate that doesn't pick out a property. But this isn't possible, at least not relative to the reasonable assumption that at least part of the meaning of a predicate is a property.

Some authors deal with this by saying that, although it doesn't pick out a property, the expression "square circle" expresses a *concept.* What is the idea here? The sentence:

(n_1) “1 is one less than the whole number that immediately precedes 3,”

doesn’t have the same meaning as the sentence:

(n_2) “1 is one less than the unique even prime.”

The property of being one less than the whole number that immediately precedes 3 is identical with the property of being one less than the unique even prime. Possession of either property is both necessary and sufficient for possession of the other. And there is no possible universe where this isn’t the case. So possession of either property is *necessarily* necessary and sufficient for possession of the other. There is nothing that distinguishes a thing’s having the one property from its having the other.

So it would be foolish to contend[] that we are dealing with two distinct properties. A better course of action would be to say that the predicates in question (“one less than the whole number that immediately precedes 3” and “one less than the unique even prime”) pick out the same property but do so by way of different *concepts*. Hence the need to distinguish between properties and concepts.

Part II Language

Chapter 3 What Is a Language?

1.0 The meaning of “meaning”

There would be no languages if there were no expressions (words, phrases, sentences, etc.). Nothing meaningless is an expression. For this reason, the concept of an expression must be understood in terms of the concept of meaning, the same therefore being true of the concept of language.

But it isn’t much use to be told that words and sentences “have meanings,” since the word “meaning” has three different meanings, and only one of these directly relates to the nature of language.

Incidentally, it’s often said that there *are* linguistic expressions that are meaningless. For example, Chomsky famously said that “colorless green dreams sleep furiously” is both meaningless *and* is a *bona fide* sentence, and therefore expression, of English.

This reasoning is no good. Depending on how it’s interpreted, “colorless green dreams sleep furiously” either says (i) that there are dreams that are green and also colorless that sleep furiously, in which case it’s false, or it says (ii) that there is nothing that is a colorless green dream that doesn’t also sleep furiously, in which case it’s vacuously true.

We’ll develop this line of thought in next chapter. In the meantime, we’ll operate on the assumption that nothing meaningless is a linguistic expression. Right now we’ll identify the three different meanings of the word “meaning” and we’ll identify the sense in which linguistic expressions are meaningful.

Meaning #1: The evidential meaning of "meaning"

In some cases, to say that *x* "means" *y* is to say that *x* is *evidence* of *y*—that *x* and *y* are causally interrelated in such a way that, given *x*, it can reasonably be inferred that *y*. "Smith's hacking cough means that he has a violent lung infection" means "Smith's hacking cough is evidence that he has a violent lung infection." And the latter means that coughs like Smith's are *causally connected* to violent lung infections in such a way that it may reasonably be inferred that Smith has a violent lung infection.

Smith's violent lung infection is a *cause* of Smith's hacking cough. But for *x* to be evidence of *y*, it is neither necessary nor sufficient that *y* cause *x*. Why isn't it *necessary?* First of all, causes can be evidence of their own effects. (Bill's current drunkenness is evidence of a poor performance on his upcoming economics exam.) Second, if some event or state of affairs *z* is a cause of both *x* and *y*, then *x* can be evidence of *y* without being a cause or an effect of *y*. (Suppose that Bill is slurring his words. This is evidence that he'll do poorly on the upcoming test. But his slurring his words is neither a cause, nor an effect, of his substandard test-performance. His drunkenness is a cause of (a) his slurred speech and (b) his imminent, substandard test-performance. And it's because his slurred speech has a causal ancestor in common with his poor test-performance that the former is evidence of the latter.) Why isn't it *sufficient?* Given only that the cause of Bill's failing was that he was drunk, we can't infer from the fact that he failed that he was drunk. There are many reasons why a person may fail a test. Supposing that *y* caused *x* to occur, *x* is evidence of *y* only if it can't reasonably be supposed that anything other than *y* was the cause. (Only a violent lung infection could be responsible for Smith's hacking cough, that being why the latter is evidence of the former.)

Meaning #2: The psychological meaning of "meaning"

When we use sentences, we mean things by them. Meaning in this sense is a psychological notion. You tell me that Sally is the most wonderful, decent person you've ever known. I respond by saying "things aren't always as they seem." What I mean is that Sally is devious. In other words, my intention in making this statement is to say that Sally is devious. Given that intentions are psychological entities, meaning in this sense is obviously a psychological notion.

Meaning #3: The linguistic meaning of "meaning"

The sentence "snow is white" says something about something; it attributes the property of being white to snow. Therefore, it *means* that snow is white. This kind of meaning is in a class by itself; it isn't identical with either of the two kinds mentioned so far; and there isn't any obvious way to understand it in terms of them. Let us now say why.

Meaning in the psychological sense *involves*, but does not coincide with, meaning in the linguistic sense. Once again suppose that, in response to your telling me that Sally is a wonderful person, I say "things aren't always as they seem." *My* meaning—what I'm trying to get across—is that Sally is devious. But in my attempt to get this across, I'm taking advantage of the fact that "things aren't always as they seem" has an existing (linguistic) meaning. My meaning that Sally is devious is parasitic on my utterance's meaning that things aren't always as they seem. Meaning in the psychological sense is therefore parasitic on meaning in the linguistic sense.

1.2 Why meaning #3 ≠ meaning #2

Some philosophers and linguists have held that for:

(1) “Snow is white,”

to mean that snow is white is for it to be the case that, in uttering “snow is white,” what *people* mean is that snow is white. This view, duly generalized, is that for a sentence *S* to have meaning *M* is for it to be the case that, in uttering *S*, people mean *M*.[135]

This position is false. There are many sentences that have determinate meanings even though they’ve never been uttered before and, therefore, no one as of yet has ever meant *anything* by them. The sentence:

(2) “The cube root of three is Sir Lawrence Olivier’s favorite irrational number between one and four,”

has a determinate meaning, and it had a determinate meaning before any part of this book was written, even though, we may reasonably suppose, it had never previously been uttered or otherwise tokened. Thus, meaning in the linguistic sense is not in *all* cases identical with meaning in the psychological sense.

But a stronger point is warranted. Let’s say that P_1 and P_2 are the propositions meant by (1) and (2), respectively. In saying that (1) means P_1 and that (2) means P_2, we are not using the word “means” equivocally. Both occurrences of “means” in the last sentence denote the same relationship. Thus, the relationship that (1) bears to its meaning is the same as the relationship that (2) bears to *its* meaning. Given that, as we saw, (2)’s having P_2 for its meaning isn’t identical with P_2’s being what people mean in uttering (2), it follows that (1)’s having P_1 for its meaning isn’t identical with P_1’s being what people mean in uttering (1). Of course, what we just said about (1) and (2) can be said of any other sentence. So even if what people mean in uttering a given sentence happens to coincide with its literal meaning, what it is for sentence *S* to have proposition *P* for its literal meaning isn’t for people to mean *P* in uttering *S*.

Psychological meaning presupposes linguistic meaning. What a person

means when uttering a given sentence is a function of, among other things, his beliefs as to what that sentence *already* means. You must believe that "snow is white" means *snow is white* if, intending to speak sincerely and literally, you say "snow is white" with the intention of getting it across that snow is white. If you think that "snow is white" means *grass is green*, you cannot, if your intention is to speak sincerely and literally, believe that "snow is white" means *snow is white.*

Of course, you could know full well what "snow is white" in fact means, but use that sentence to get across something that has nothing to do with the color of snow. Knowing what "snow is white" actually means, you might utter that sentence with the intention of getting it across that the government is controlling our thoughts with alpha waves. And, depending on the circumstances, that could be precisely what an utterance of that sentence would convey.

But whatever the message is that you wish to convey in uttering a certain sentence, you must believe that message to have *some* kind of relationship to the one meant by that sentence itself. Furthermore, if you are to *succeed* in saying what it is you wish to say, what you believe to be meant by the sentence you are using must be *right.* If, intending to speak sincerely and literally, you say "snow is white," thinking that it means *bananas are yellow*, you will fail to say what you wanted to say.

Thus, setting aside defective utterances, one cannot, in uttering a given sentence, mean *anything* by it unless one knows what it *already* means. So meaning in the *psychological* sense is parasitic on meaning in the linguistic sense, and the two kinds of meaning are therefore entirely distinct.

1.3 Why meaning #3 ≠ meaning #1

The sense in which "snow is white" means that snow is white isn't comparable to the sense in which smoke means fire. The fact that smoke means fire has nothing to do with conventions on the part of people.[But the fact that "snow is white" means what it does *is*, at least in part, a matter of convention. It's a matter of convention that "snow" doesn't refer to grass and, therefore, that "snow is white" doesn't mean that grass is white; it's a matter of convention that "white" doesn't mean *green* and, therefore, that "snow is white" doesn't mean that snow is green.

Although the whiteness of snow sometimes causes people to say "snow is white," it doesn't do so in the way that fire causes there to be smoke. Fire happens; smoke happens as a result. The presence of smoke doesn't embody any *judgment* about anything. But, when caused by the whiteness of snow, utterances of "snow is white" *do* embody judgments of various kinds. People see or otherwise come to believe that snow is white; and, since they know the relevant linguistic rules, they know that were they to say "snow is white," they judge that they'd be making a *correct* statement. Thus, utterances of "snow is white" embody judgments about the color of snow and about how, linguistic conventions being what they are, one can report the color of snow. Also, people don't say everything that occurs to them. Before deciding to utter a given sentence, people typically make context-based judgments about the appropriateness of uttering that sentence. So various judgments—about snow, about language, and about human psychology—are involved in the causal connection between the whiteness of snow and a given person's saying "snow is white." There is thus a *normative* dimension to language use that is absent where purely natural, non-conventional cause-effect relations are concerned.

2.0 Sentences as proposition-isomorphs

The meaning of a true or false sentence is a ***proposition.*** Propositions are not themselves sentences. That is why different sentences (e.g., “schnee ist weiss” and “snow is white”) can express the same proposition.

Propositions, when true, are *truths.* Thus, propositions have existed as long as there have been truths; which means that they’ve existed as long as there has been anything and, consequently, that propositions are not our creations.

Though distinct from the sentences that express them, propositions are structurally similar to them. Two otherwise dissimilar sentences can share the word “John.” “John loves Mary” and “Sally punched John” are two such sentences. The meanings of those sentences obviously have something in common corresponding to the fact that they share the word “John.” Since they share no other constituents, the thing meant by “John” must be capable of moving *on its own* from sentence-meaning to sentence-meaning. This would not be the case if the thing meant by “John” in the proposition meant by “John loves Mary” were incapable of being disengaged from the things meant by “loves” and “Mary.” It follows that propositions consist of discrete parts; it also follows that those discrete parts correspond to discrete parts of the sentences that express them. Taken together, these two points entail that sentences are structurally like the propositions they express.

2.1 Propositions as digital structures

Given that propositions consist of discrete, isolable entities, it follows that, like sentences and unlike visual perceptions and photographs, propositions are *digital* structures. The sentence "Sally punched Bob" has a unique decomposition into a certain "minimal units of significance," or "morphemes," these being "Sally," "Bob," etc. Given what we said in Section 2.0, it follows that something similar is true of the corresponding proposition. Sentences and propositions are *digital* structures, meaning that they have unique breakdowns into minimal significant units.

A visual perception of Sally punching Bob doesn't have a structure comparable to that of "Sally punched Bob" or any other sentence. Unlike sentences, perceptions don't have to decompose into minimal significant units. Visual perceptions, unlike sentences and propositions, therefore have non-digital or *analogue* structures. Given that at least some thought involves the processing of perceptual information, it follows that thought

at least sometimes has a structure very different from language. (See Section 5.4 for further discussion of this.) The nature of propositions is discussed in Chapter 4.

3.0 The three branches of the philosophy of language: semantics, pragmatics, and syntax

The study of language is typically divided into three sub-disciplines—semantics, pragmatics, and syntax. In addition to denoting a branch of linguistic study, each of these three words denotes dimension of language. So “semantics” refers to a certain discipline and also to a feature of expressions, the same being true of the other two expressions.

3.1. Semantics

The discipline of semantics attempts to make it clear what our utterances literally mean. It has no interest in what is conveyed through suggestion or innuendo.

If a disappointed boss says to a substandard employee, “it might not be a bad idea for you to start thinking about finding a new position,” the literal meaning of his utterance is quite innocuous. But the message that is being *sent* is not innocuous—that message is: *you’re fired; you’re a disgrace; go away;* etc. That utterance's *semantics* coincides with its literal (innocuous) meaning.

3.2 Pragmatics

Pragmatics studies the *use* of language. Sometimes language is used literally. Asked whether I'm over thirty years of age, I say "yes, I'm over thirty years of age." What *I* mean coincides with what my utterance literally means.

Language is often used non-literally. If, while addressing a pan-handler, I say "you've made a fine life for yourself," what I mean is the antithesis of what my utterance means. But usually the propositions literally meant by our utterances are neither opposed to, nor exhaustive of, the propositions we wish to affirm in producing those utterances. Asked whether there's a place to get food, I say "there's a McDonalds down the road." The proposition I'm affirming is: there is a nearby place to get food, the reason being that there's a McDonalds down the road. Thus, the proposition literally meant by my utterance is only a part of what it is that I'm saying. Thus, what a sentence literally means is only one of many factors governing what it is used to mean. The discipline of pragmatics tries to identify the remaining factors.

3.3 Syntax

The discipline of syntax studies the structures of the meanings of ***complex expressions.*** A complex expression is one that consists of other expressions. (Thus, “the man who ate my cookie” is a complex expression, since it consists of “man,” “ate,” etc., each of which is meaningful. By contrast, “red” is not a complex expression, since it doesn’t have any meaningful proper parts.) The discipline of syntax tries to make it clear how the meanings of complex expressions depend on those of their parts.

Thus, the discipline of syntax doesn’t study the meanings of complex expressions *per se.* It studies the *relationships* that such meanings have to those of their constituents. Consider the sentence “Sally hates Bob.” The word “hates” ***occurs*** in that sentence. If that occurrence is replaced with an occurrence of “loves” or “is amused by,” the resulting sentence has a very different meaning from the first. This shows that what “Sally hates Bob” means depends on what “hates” means.

Bearing this point in mind, consider the sentence “Larry loves Julie.” Obviously this sentence doesn’t mean the same thing as “Sally hates Bob.” But the *relationship* borne by the meaning of “Sally hates Bob” to the meaning “hates” is identical with the relationship borne by the meaning of “Larry loves Julie” to that of “loves.” Exactly similar points hold in connection with each of the remaining two constituents of each of those sentences.

The discipline of syntax studies the *relation* that the meanings of complex expressions bear to the meanings of the simple expressions composing them. Thus, syntax doesn’t study the semantics (meanings) of complex expressions. It studies the *structures* of the semantics of complex expressions. Syntax studies semantic structure.

4.0 The need for the discipline of semantics

Even though we all know what is meant by:

(1) “John wants to catch a 20-pound striped bass,”

we don’t know what it is that we know in knowing this. Semanticists supply us with the missing meta-knowledge. Consider the sentence:

(2) “John wants to punch Bob.”

(2) attributes a certain property to John—that of wanting to punch Bob. Given that (1) and (2) are grammatically isomorphic, it’s natural to assume that there exists some 20-pound striped bass *x* such that the proposition expressed by (1) is:

(3) John wants to catch *x*.

But this isn’t the right analysis. There isn’t some *one* fish such that, if the desire ascribed to John by (1) is to be gratified, John must catch *that* very fish. There is thus no fish *x* such that, if (1) is to be true, John must want to catch *x*.

The meaning of (1) is:

(1_R)[138] John wants it to be the case that: *there exists some fish x such that x is a 20-pound striped bass and such that John catches x.*

Thus, (1) doesn’t describe a relationship between John and some non-existent or quasi-existent fish. It affirms the existence of a relationship between John and a *proposition.* The proposition in question is one that, in English, is expressed by the sentence:

(4) “There exists some fish *x* such that *x* is a 20-pound striped bass and such that John catches *x*.”

If John’s fishing-trip is a success, that proposition will be true; otherwise it

will be false. But that proposition exists either way.

But we still haven't solved the problem. In (4), the expression "some fish" occurs. Grammatically, that expression is a noun-phrase. But, unlike other noun-phrases, it doesn't refer to anything. "Some fish" doesn't refer to some fish. There is no fish *x* such that "some fish" refers to *x*. That's why, given any particular fish *F*, if you say ‹some fish is wet, but *F* is not›, what you are saying isn't self-contradictory. So the problem we were trying to solve remains.

But to solve the problem, we need only reword (4). The needed rewording is this:

> (5) "The property of being a 20-pound striped bass that John catches is instantiated."

The property of being such a fish exists. So (1), which seemed to be about a non-existent fish, is about an existent property. (5) says of that property that it's instantiated. Thus, a complete analysis of (1) is given by

> (1_{CA}) John wants it to be the case that the property of being a 20-pound striped bass that John catches is instantiated.

So even though just about every English speaker understands (1), knowing what it is that one knows in understanding it isn't such a trivial thing.

4.1 Semantics is needed to figure out what is literally meant and what is not

Despite everything just said, there is clearly a sense in which every English speaker knows what (1) means. What the semanticist is doing in connection with (1) isn't comparable to what somebody who speaks both Spanish and English is doing in connection with it when he tells his monolingual Spanish speaking friend what it means. The semanticist is needed to clarify the structure of the meaning that (1) is already known to have, but he isn't needed to *identify* that meaning. The semanticist isn't a translator. But there are many cases where the semanticist *is* needed to identify literal meaning. In fact, as paradoxical as it may sound, there are cases where he is needed to identify the meanings of sentences *that are perfectly well understood.*

First of all, we must distinguish what is literally meant by an utterance from what it is that the speaker wishes to convey. To give a trivial example: You and I are robbing a bank. I yell: "The cops are coming!" What I wish to convey is that we should hurry up. In this particular case, it's easy to distinguish what is literally meant from what is non-literally suggested, and semantics would therefore have no interest in it. But in other cases, it's exceedingly hard to do this, and it's with these other cases that semantics is concerned.

A story will help us move forward. Somebody who is wearing a ski-mask, and who I therefore don't recognize, deftly snatches my wallet from my pocket. As he's running off, I point at him and yell: "That man is a thief!" Let U_1 be this utterance.

Before moving on, let's take a moment to make it clear what U_1's literal meaning is. Somebody just stole my wallet. I don't know who that person is. But whoever it is, I am attributing a certain property to him. If that person has that property, I have spoken truly; if not, I have not. U_1 is correct if, and only if, the person referred to by "that man" has the property expressed by "is a thief." Thus, there is some individual *x* such that *x* has just stolen my wallet and such that what I've just said is true exactly if *x is a thief.* (The italicized part is U_1's literal meaning.)

The next day, my lovable office-mate Steve eats one of the cupcakes that was on my desk. I jokingly point at him and say: "That man is a thief." Let

U_2 be *this* utterance. There is some *x* such that *x* just ate my cupcake and such that U_2 is true exactly if: *x* is a thief.

Unbeknownst to me, Steve is the pick-pocket, and there is some individual *x*, namely Steve, such that each of U_1 and U_2 is true if and only if *x* is a thief. Thus, U_1 and U_2 have the very same literal meanings.[140] But I don't know this, even though I speak English perfectly and, on each occasion, obviously understand perfectly well what it is that I'm saying.

How is this possible?[141] Our sense-perceptions describe things. My uttering U_1 was a response to my being given a visual description of Steve. That description was to the effect that:

> (i) There is some man *x* such that *x* is wearing a ski-mask and such that *x* is running off into the distance.

But U_1's literal meaning is not to that that effect. There is some man *x* such that *x* is a wearing a ski-mask (etc.), such that in uttering U_1 I was saying that:

> (ii) *x* is a thief.

The meaning of U_1, being identical with (ii), is quite threadbare. But I grasped that threadbare meaning through my descriptively rich visual perception, whose content is given by (i). My uttering U_2 was a response to my being given a different description of Steve. That description was to the effect that:

> (iii) There is some man *x* such that *x* is a portly amicable fellow who is sitting over in that chair.

But U_2's literal meaning is not that that effect. There is some man *x* such that *x* is a portly amicable fellow (etc.) such that in uttering U_2 I was saying that:

> (ii) *x* is a thief.

Echoing what we said a moment ago, the meaning of U_2, being identical with (ii), is quite threadbare. But I *grasped* that threadbare meaning through my descriptively rich visual perception, whose content is given by (iii).

Because I grasped (ii) by way of different bodies of perceptual (descriptive) information, I didn't know, when uttering U_2, that what I was affirming was the same thing I was affirming in uttering U_1. Oftentimes, literal meaning is cloaked by the *pre*-semantic information through which it is grasped, and semanticists are needed to uncloak it.

4.2 Semantics are needed to figure out what is literally meant and what is not (continued)

Fido is the smartest dog on the planet. I know this well, but some of my friends don't yet know this. I point to Fido and say: "That dog is very smart."

The proposition that it was my intention to affirm and communicate is indeed true. *For a dog,* Fido is indeed smart. Of course, Fido is vastly less intelligent than a human being, such as my friend Timmy, who is of mediocre intelligence. But if I say "Timmy is very smart," what I'm saying is *false.* Judging by the words I've used, the property I've attributed to Timmy is identical with the property I've attributed to Fido. Given that Timmy has that property to a vastly greater degree than Fido, it would seem to follow that, since "Fido is smart" is true, "Timmy is smart" must also be true. And yet "Timmy is smart" is false. How can this be?

Some deal with this by saying that "smart" is ambiguous, like the word "dumb." So "Timmy is smart" and "Fido is smart" have different meanings, like "Timmy is dumb [unintelligent]" and "Timmy is dumb [mute]."

This solution is pretty clearly false. A more plausible one is that the property of being smart *for a dog* is distinct from the property of being smart *for a human.* Fido has the first but not the second. And many humans have the second, but almost all of those lack the first.

A similar, possibly coincident, view is that "smart" is implicitly relational. When, for some object *x*, you say ‹*x* is smart›, you are saying that *x* is smart *relative to some benchmark,* the identity of which the context makes clear. So "Fido is smart" says that Fido is smarter than most dogs, which is true, and "Timmy is smart" says that Timmy is smarter than most human beings, which is false.

I published a paper[142] arguing that, for any degree-property phi, ‹*x* has phi› expresses a proposition of the form: the degree to which *x* has phi exceeds standard *S*, where *S* is some standard that, given the context, is clearly the relevant one. (A "degree property" is one that can be had to varying degrees.) But even if this is right, it doesn't follow that such judgments are the literal meanings of such sentences. And there is no independent evidence that ‹*x* is smart› has the syntactic properties of sentences that clearly *do* have for their literal meanings propositions of the just-described kind. This suggests that, so far as ‹*x* has phi› communicates such a proposition, it isn't because it semantically encodes it.

In any case, it is not *obvious* what ‹*x* is smart› means or, in general, what ‹*x* has phi› means, where phi is any degree property. Thus, the literal meanings of such sentences are sufficiently recondite that the intervention of professional semanticists is needed to identify them.

See Chapter 4 for a thorough discussion of the points just made.

5.0 Semantic rules

The English language assigns a certain meaning to the sound “that dog has rabies”; and given the spectacle of a rabid dog, it furnishes one with a sentence with which to describe what one sees. In general, the English language assigns meanings to sentences and sentences to meanings. This is true of all languages. A language is a systematic way of pairing off sentences with meanings. Any rule that assigns a meaning to an expression is known as a “semantic rule.” Languages are sets of semantic rules.

5.1. An important subtlety

There are a couple of subtle but, in some contexts, important inaccuracies in what I just said. First of all, something isn't a sentence *until* a meaning has been assigned to it. In a world where there were no animate beings, but in which the forthcoming parenthetical ink deposit (1 + 1 = 2) was formed out of twigs, that twig deposit wouldn't be an expression of any kind. That twig deposit *would* be an expression *if* and *only if* it were endowed with meaning. This shows that something has to have meaning *in order* to be an expression. Thus, *expressions* aren't assigned meanings. They already have them and don't need to be assigned them. Therefore a semantic rule can't be defined as a rule that assigns a meaning to an expression.

5.2 How meaning is assigned to hitherto meaningless, and, therefore non-linguistic, entities

Thus, semantic rules assign meanings to non-expressions. But which non-expressions? A story will give us the answer.

You and I want to invent a code that only we two know. We both know a guy whose real is name is Larry. We decide that our code name for Larry is to be "Ichabod." So what's going on is that we're creating a semantic rule: one that to the effect that "Ichabod" is to pick out Larry. How exactly is this rule enacted?

In order to implement this rule, I say: Let's refer to Larry as "Ichabod." The burst of noise that I produce is a sentence-token. And what you *hear* is some *token* of "Ichabod"—you do not, since one could not, hear the name type itself.

When you hear this burst of noise, along with my proposal concerning our new name for Larry, you know that what I'm saying is to the effect that any other physical object that is similar in the relevant ways to this burst of noise is itself henceforth to refer to Larry. Thus, I am in effect proposing that all and only those bursts of noise that are similar, in the relevant respect, to this particular burst of noise are to refer to Larry. (By implication, I'm proposing the same thing *mutatis mutandis* to hold of all and only ink-deposits that, given certain conventions, are paired off with such bursts of noise.) The thing that, according to my proposal, is henceforth to pick out Larry is the thing of which all and only such bursts of noises (etc.) are *instances.* That thing, like

anything else of which there are instances, is a property. It is the property had in common by all and only bursts of noise (etc.) of the relevant type.

That property doesn't (yet) have a meaning; it isn't (yet) an expression. It's a property that existed, and was instantiated, before either or any instances meant anything. So the semantic rule that I'm proposing we adopt assigns a referent to a property that does not itself have a meaning. The same thing *mutatis mutandis* holds of any other semantic rule.

Thus semantic rules assign meanings to properties of physical objects—to morphological or acoustical properties (in other words, to properties that a thing has in virtue of having a certain shape or sounding a certain way). A semantic rule is therefore something which assigns a meaning to a property, and language is a set of such rules.

(Technically, this is only an approximation to the truth. The relevant qualifications are found in Section 7.5.)

5.3 Semantic rules (revisited)

Many believe that linguistic meaning is to be understood in function-theoretic terms. More precisely, they believe that semantic rules are mathematical ***functions***.

Let us start by defining the word "function." Given any pair of whole numbers, the expression "+" assigns exactly one whole number to that pair. Thus, "+" denotes a function that assigns single numbers to pairs of numbers. In general, a function is a rule that, given some class of objects, assigns no more than one object to any given member of that class.

Although the rule expressed by "plus" assigns the number 8 to the pair <4,4>, it doesn't assign that number to that pair in the way in which a person assigns a task to an underling. In the one case, the word "assigns" has a psychological meaning and it refers to something spatiotemporal; in the other case, it has a non-psychological, purely logical meaning, and it refers to something non-spatiotemporal. A related point is that the rule that assigns 8 to <4,4> isn't a social rule, and it therefore isn't something that can be obeyed or disobeyed.

According to the function-theoretic view, semantic rules are rules in the same strictly logical sense in which $F(x)=x+2$ is function. The semantic rule for "Socrates" is a function that assigns a certain individual (Socrates) to that word (or to occurrences thereof). The semantic rule for "snow is white" is a function that assigns truth-conditions to that utterance (or to occurrences thereof). And so on.[145]

5.3.1 Why semantic rules are not functions

The just-described view is false. The rule denoted by "+" has always existed and always will. Of course, the *expression* "+" hasn't always existed. But that's irrelevant, since things pre-exist the expressions we use to denote them. "Socrates" is an Anglicization of the name by which Socrates was referred to by his contemporaries. Since Socrates lived well before the English language came into existence, "Socrates" (the name, not the person) didn't come into existence until well after its referent went out of it.

The semantic rule that assigns Socrates to ink deposits having certain shapes would exist even if the English language had never come into existence. Like the rule denoted by "+" that rule has always existed and always will. So has the rule that assigns Abraham Lincoln to such ink deposits. The mathematical function that assigns the proposition *snow is white* to ink deposits like the italicized one has always existed, as has the mathematical function that assigns the proposition *all horses weigh 18,000 lbs.* to those same ink deposits.

But the English language hasn't always existed. Since the English language *is* a set of semantic rules, those semantic rules haven't always existed. They came into existence quite recently and thus cannot be mathematical functions.

Also, if the semantic rules of English *were* such functions, there would exist a language in which "Socrates" referred to Lincoln and in which "snow is white" meant *all horses weigh 18,000 lbs.*, the reason being that the corresponding mathematical functions exist. But there is no such language. Of course, there *could* be such a language. And maybe there will be; maybe somebody will invent a code in which those things have those meanings. But right now they don't. Such a language is merely possible, and therefore doesn't exist. Thus, semantic rules are not rules in the mathematical sense.

5.3.2 The Gricean approach

Understandably, many philosophers of language believe that semantic rules must be understood in psychological, not mathematical, terms. There are different versions of this view. I accept one version of it. But the version I accept bears little resemblance to the versions of it that are usually held, each of which is some variant of the view held by H.P. Grice.

According to Grice (1957), for expression E to have literal meaning M is for it to be the case that, when they utter E, M is what they mean. So their answer to the question "why does 'snow is white' have the proposition that snow is white for its literal meaning?" is "because what people generally mean when they say 'snow is white' is that snow is white." Aware of some of the problems with Grice's view that we're about to identify, neo-Griceans concede that, although literal meaning cannot in all cases be *identified* with speaker's meaning, the former is parasitic on the latter in ways that are correctly described by modest liberalizations of Grice's view.

Wittgenstein (1958) advocated a version of this view. "Roughly speaking," he said, "meaning is use." Expressions mean what we use them to mean—they mean what *we* mean by them. Wittgenstein nowhere makes it clear what he means by the words "roughly speaking."

But it doesn't matter, since his statement isn't even roughly true. Literal meaning isn't identical with speaker's meaning and it isn't to be understood in terms of it. It's the other way around. We saw why in Section 1.3.

Also, Grice's theory fails to deal with the fact that the meaning of a subsentential expression isn't something that could possibly be meant. "Of " has a meaning; so does "skip," "snorkel," "or," "gladly," etc. But whatever it is that "or" means, it cannot, at least not by itself, be what a *person* means by it. I cannot say "or" and *mean* it---my meaning cannot be its meaning.

5.3.3 More problems with the Gricean approach

If Grice were right, the meaning of the sentence:

(SF) “Smith is now living in France,”

would be fixed by the intentions people have in using it.

But its meaning is *not* fixed by those intentions. It is fixed by the meanings of its parts (“Smith,” “France,” etc.), together with the way those expressions are ordered in that sentence. The semantic and syntactic rules of English being what they are, SF would have its current meaning even if it had never been used. So whatever the intentions of people using that sentence are, those intentions do nothing in the way of assigning it that meaning. In general, Grice’s view is inherently incapable of accommodating the fact meaning is compositional.

5.3.4 How some Griceans deal with the problem just described

Some Griceans respond by saying that[146]: [H5]

> (GV) [147] Although speaker-meaning doesn’t *directly* fix sentence-meaning, it does so indirectly. In uttering sentences of the form ‹....France...›. our intention is to say that…*France*… (For example, in saying "France is lovely,” one's intention is to say that France is lovely; and in saying "Jerry loves vacationing in France", one's intention is to say that Jerry loves vacationing in France.) It is in virtue of that fact that “France” refers to France.
>
> For “France” to refer to France *is* for ‹....France...› to mean that…*France*…So even though our linguistic intentions don’t directly endow “France” with meaning, they do directly endow sentences containing “France” with meaning, and they thus *in*directly endow “France” with meaning. Therefore, Grice’s core idea is right: What an expression means is to be understood in terms of what people using it are trying to say. [148]

GV is a failure. Let P be the proposition meant by SF. So far as people utter SF with the intention of affirming P, it's because they believe (correctly, as it happens) that each of the expressions composing it already has a certain meaning.

It's irrelevant that how we use sentences of the form ‹...France...› *causally* determines what "France" means. There are many ways to *cause* meaning-shifts—many ways to *get* a given expression to have a certain meaning. But the question we're asking isn't "how did 'France' *acquire* its current meaning?", and is instead "whatever it is that 'France' means, what *is* it for it to have that meaning?" And it's no answer to this question to say that it may have acquired that meaning because of what, at some point in time, people meant by sentences of the form ‹...France...›.

5.3.5 Why Grice's theory is inconsistent with the normative nature of semantic rules

A billiard ball isn't right to move after being struck; it just does. The relevant scientific laws merely register that fact; they aren't normative—that is, they don't characterize it as good or bad. Unlike scientific laws, semantic rules *are* normative. If, intending to affirm that Smith is female, you say "Smith is male," you've done something wrong. If Grice is right, literal meaning is speaker's meaning. This means that, if Grice were right, speaker's meaning wouldn't be accountable to existing semantic rules. Since it is, Grice is wrong.

5.4 The psychological reality of semantic rules

One view as to the nature of semantic rules is that they are idealized descriptions of the activities of speakers. Proponents of this view seldom if ever identify the facts about the speaker-behavior of which semantic rules are supposedly descriptions. The most natural assumption is that they are idealized descriptions of what people *mean* when they speak and write (etc.). If this assumption is right, then, given the points just made, the view in question is wrong.

But even if this isn't what proponents of this view have in mind, their view is very clearly wrong. If semantic rules are just idealizations of speaker-behavior, then speaker-behavior must pre-exist the semantic rules embodied in it. But if that's the case, then the activity described by semantic rules isn't *guided* by them. An awareness of those rules is no part of what leads people to say the things they do. Those rules are psychologically inert. They have no "psychological reality." This view is held by Nathan Salmon (2007) and also by Scott Soames (2002).

This view is inconsistent with some obvious facts. I know that Smith is now living in France. Wanting to tell you this, I say: "Smith is now living in France." Why do I choose this particular sentence? Because, first of all, I know the relevant semantic rules (viz., that "Smith" refers to Smith,that "living" refers to a certain property) and, secondly, because I believe that, given these facts about semantics, the sentence in question is the right one to express my belief. A knowledge of semantics underlies my speech-act and, by obvious extensions of these points, all non-defective speech-acts.

Another problem with the Salmon-Soames view is that it's inconsistent with the normative character of semantic rules. If semantic rules merely describe existing semantic activity, then that activity isn't answerable to semantic norms. Semantic activity obviously is answerable to such rules. If, intending to tell you to take a break after jogging for another 100 yards, I say "jog continuously for another ten miles; therefore, do not take a break after jogging for another 100 yards," I've misspoken, given what it was that I intended to say, and I've therefore violated the relevant semantic norms.) So the Salmon-Soames view is incorrect.

5.5 Conceptual role semantics

A little while ago, we briefly discussed Wittgenstein's claim that "meaning is use"; i.e., that for an expression to have a given meaning *is* for it to be used in a certain way. This doctrine is incoherent in many ways. We've already discussed one of those ways; now we'll discuss some of the others.

Expressions have meanings. A meaningless burst of noise isn't an expression. If I cough or guffaw, the burst of noise I've produced doesn't have the sort of meaning had by *bona fide*expressions. It has, at most, meaning in irrelevant, purely evidential sense; e.g., the sense in which a cough may be evidence of a cold. Since anything that is a linguistic expression *ipso facto* has a meaning (in the relevant, linguistic sense), there are no expressions to be used *before* noises, ink-marks, etc., have been assigned meanings. So, since there can be no expression-use until after there is expression-meaning, it makes no sense at all to say that expression-use determines expression-meaning. "Meaningful expression" is a pleonasm.[149] So, contrary to what Wittgenstein said, meaning isn't use.

Of course, how a given expression is used may well assign it a *new* meaning. But there is all the difference in the world between saying:

> (1) Expression *E*'s having meaning *M* is *causally* determined by *E*'s being used in such and such a manner,

and:

> (2) Expression *E*'s having meaning *M* is *identical with E*'s being used in such and such a manner.

(1) is a sterile triviality. And (2) is wrong. An expression *E*'s having meaning *M* cannot possibly be *constituted* by its being used in such and such a manner since *E* isn't an expression, and therefore isn't an expression to be used *until* it has a meaning.

According to a doctrine known as "conceptual role semantics" (CRS), whose exponents include Hartry Field (1977) and Robert Brandom (1994), for an expression to have a given meaning is simply for it to be used in a certain way. But this isn't correct as we just saw.

CRS is incoherent for reasons other than the one just given. According to that doctrine, what a sentence means is determined by what people infer from it and what people infer it from. Whereas commonsense holds that one infers "an even number is less than five" from "two is less than five" *because* the latter already has a given meaning, advocates of CRS say that, on the contrary, it's because people infer "an even number is less than five" and other similar statements from "two is less than five" that the latter has the meaning it has.

This is not a viable view. If we learn tomorrow that, contrary to what we previously thought, Aristotle wrote several plays, we'll infer "somebody who wrote several plays wrote the *Nichomachean Ethics* from "Aristotle wrote the *Nichomachean Ethics*." But it doesn't follow that "Aristotle wrote the *Nichomachean Ethics*" would have undergone some change in its semantic *meaning.* Changes in what we believe affect what we infer from statements; but they don't categorically change the meanings of those statements. CRS entails that every inference is an analytic inference. Once it's learned that Aristotle wrote plays, it becomes, according to CRS, constitutive of the *meaning* of "Aristotle wrote the *Nichomachean Ethics*" that one can infer from it that a playwright wrote the *Nichomachean Ethics.* But surely the inference from "Aristotle wrote the *Nichomachean Ethics*" to "a playwright wrote the *Nichomachean Ethics*" isn't *analytic*.[150]

What we may infer from a sentence is answerable to its existing meaning. CRS says that a sentence's meaning is answerable to what we infer from it. If correct, that would have the consequence that one couldn't possibly draw a false inference from any sentence. This in its turn would have the consequence that no sentence would mean anything. Which, since nothing meaningless is a sentence, would have the absurd consequence that there neither are, nor could be, sentences.

Consider the sentence "Bill plagiarized his first novel." If one knows that sentence to be true, one can make inferences about Bill's character, his past activities, his ambitions, his values, and so on, that one couldn't make if, other things being equal, one didn't know that sentence to be true. But it's only because of what the sentence *already* means that one can make those inferences. Given any other sentence, the same thing *mutatis mutandis* is true of it. CRS says that what a sentence means is determined by what we infer from it. Since it's the other way around, as we've just seen, CRS is false.

6.0 What is literal meaning?

Where complex expressions are concerned, there is no limit to how much literal and understood meaning may diverge from each other. But where simple expressions are concerned, literal and understood meaning must coalesce. It makes no sense to suppose that people could be systematically wrong as to what "red" meant. If people *thought* that "red" meant what is in fact meant by "blue," then "red" *would* have that meaning. Systematic, widespread error is impossible where semantically simple expressions are concerned. This gives us a way of understanding what literal meaning is.

Even though there can be widespread, systematic misinterpretations of sentences, those misinterpretations do not arise as a result of people failing to know what the simple parts of sentences mean. They arise as a result of people not knowing how to put those meanings together. So *to the extent* that its meaning is fixed by the fact that it has the form ‹...Socrates...›, people (English speakers) *do* systematically understand "Socrates was more wise than Plato, but he was less sharp then Aristotle"; and *to the extent* that its meaning was determined by its having the form ‹...wise...›, people *do* understand that sentence; and so on. So far as that sentence *is* systematically misunderstood, it is because people are having trouble putting the meanings of its constituents together—it is because they're having trouble figuring out how those meanings ought to be put together.

What a simple expression literally means is determined by what it is that a sentence means by virtue of containing it. Since, where simple expressions are concerned, what people take literal meaning to be coincides with what it really is, a simple expression's literal meaning coincides with what it is that, in virtue of containing that expression, sentences are *taken* to mean. So a simple expression's literal meaning is given by a statement saying what it is that, by virtue of containing it, sentences are taken to mean; and a complex expression's literal meaning is a function, in the mathematical sense, of the meanings of its parts.

Here's an illustration. "Socrates" is a simple expression. So what people think it means must ultimately coincide with what it actually means. So it refers to Socrates only because people think it refers to Socrates and, therefore, only because people think that "Socrates is intelligent" attributes intelligence to Socrates and, in general, that ‹...Socrates...› attributes...*x*...to Socrates.

Of course, people don't always take utterances of ‹...Socrates... › to be

attempts to *say* that Socrates has...*x*... It might be clear from the speaker's tone that what he *really* meant when in saying "Socrates was wise" was that Socrates was *not* wise. But *to the extent* that their belief that "Socrates" refers to Socrates is determinative of what people take the meanings of utterances of the form ‹...Socrates...› to be saying, what they take it to be saying is that Socrates has...*x*... That is what it is for them to take "Socrates" to refer to Socrates. And their taking "Socrates" to refer to Socrates *is* for "Socrates" to refer to Socrates, given that "Socrates," being a simple expression, has the semantics that people think it has. The same thing *mutatis mutandis* is true of every other simple expression.

Bearing these points in mind, let CE be any complex expression, and let $e_1...e_n$ be the simple expressions composing it. How people interpret CE may diverge from its literal meaning. But when this happens, it's because what it is taken to mean diverges from what people are *disposed* to take it to mean, given what people believe its simple parts to literally mean.

Since what people take simple expressions to mean *is* what they mean, this is the same as saying the following. A divergence between:

(i) A sentence's literal meaning,

and:

(ii) That sentence's understood meaning,

is the same thing as a divergence between:

(a) That sentence's literal meaning,

and:

(b) What it is that, given their (correct) beliefs as to what its simple parts literally mean, people are *disposed* to take that sentence's literal meaning to be.

The literal meaning of a complex expression is a function of two things: (i) the meanings of its simple parts, and (ii) the order in which those parts are arranged. To say what literal meaning is in general, we need to say what it is

for a simple expression to have a given literal meaning. Given the points just made, we can do this. Where simple expressions are concerned, literal and communicated meaning coincide (ultimately)—in other words, such expressions mean what people take them to mean. And where complex expressions are concerned, literal and communicated meaning may diverge, but literal meaning nonetheless coincides with what people are *disposed* to take it to mean, given what the simple components of the expression in question literally mean.

7.0 Tokens vs. types: Some preliminary terminological points

No word is identical with any utterance of it. My utterance of the word "snow" lasts for a fraction of a second. But the word "snow" *itself* endures.

Utterances and inscriptions of expressions are referred to as "expression-tokens" (or just "tokens"). So there are three *tokens* of some one word to the right of the upcoming colon: snow, snow, snow. The things being uttered or inscribed are referred to as "expression-*types*" (or just "types").

7.1 Two-dimensional semantics[151]

Some expressions appear to have a two-tiered semantic structure. For example, an occurrence of the pronoun "I" has a referent, this being the person who uttered it, and it picks out that referent in a certain way (i.e., by way of a certain concept).

Thus, the semantics of "I" is given by the rule that an utterance of "I" refers to a given person if and only if that person has the property of being the one that produced that utterance.

So, if I say "I am tired," the concept through which my utterance of "I" refers to me is the concept person who produced the utterance in question. This, then, is the concept through which reference is secured; it is, as we'll henceforth put it, the mediating concept.[152] When Smith says "I am tired," his utterance of "I" refers to himself, not to me. But the mediating concept remains the same.

If somebody points to me and says "that guy is tall," the person picked out by "that guy" is me. But in this case, the mediating-concept is different. The semantics rule for "that guy" is given by the rule that, if "that guy" is uttered in a context where there is a unique, salient guy, that utterance refers to that individual.

"I" and "that guy" are context-sensitive expressions: What such an expression refers to depends in a systematic manner on facts about the context of utterance. (We'll refine this vague statement shortly.) Such expressions are known as "indexicals." Other examples of indexicals are "you," "he," "those animals," "this monkey," "tomorrow," "yesterday." Some indexicals are single words (e.g., "tomorrow," "he"); others consist of more than one (e.g., "that tall man"). The latter are known as "complex

indexicals.”

7.1.1 Demonstratives vs. indexicals

Some indexicals often cannot be successfully used without an accompanying demonstration on the speaker's part, the purpose of which is to eliminate any doubt as to what the intended referent is. If Jim and Larry are both equally salient in the context in question, an utterance of "that guy" won't single anyone out. But it will do so if an act of pointing accompanies it. Indexicals that fall into this category are known as "demonstratives." Not all indexicals are demonstratives. For example, "tomorrow," "now," and "here," aren't demonstratives. Given an utterance of "tomorrow," no gesture is needed to make it clear what the intended referent is. If I say "tomorrow I'm going hiking," it is clear what the referent of "tomorrow" is; no demonstrative act is necessary, and none could possibly do any good.

7.1.2 Indexicals (continued)

Indexicals have a “two-dimensional” semantic structure. Consider the expression “today.” If I say it right now, that utterance will refer to April, 27, 2009, since that is today’s date. If I say it tomorrow, that utterance will refer to April 28, 2009. But the *rule* that assigns April 27, 2009, to the first utterance is identical with the *rule* that assigns April 28, 2009, to the second utterance. That rule is this: if “today” is uttered on a given day *D*, that utterance refers to *D*.

The meaning of an *utterance* “today” is the day it picks out. The meaning of the corresponding word-*type* is the rule just described. Given any indexical, the meaning of a *token* of that indexical is its referent; the meaning of the corresponding indexical-type is the rule that assigns that referent to that token.

7.2 Definite descriptions

If our pretheoretic intuitions are correct, definite descriptions are devices of reference.[153] It's very hard to believe that "the whole number that comes right after one" doesn't refer to the number two. In this section we will operate on the assumption that, indeed, definite descriptions are devices of reference. (Later, in Section 9.0, we'll see good reasons to reject this way of understanding definite descriptions. But, for now, we'll operate on the plausible and, by the laity, universally held believe that they are what they are appear to be----that "the number two" refers to 2, that "the inventor of bifocals" refers to Benjamin Franklin.) [154]

Definite descriptions, like indexicals, have a two-dimensional semantic structure. A given utterance of "the current U.S. President" refers to some individual. Right now such an utterance would refer to Barack Obama. A few years ago, such an utterance referred to Bill Clinton.

But even though an utterance in 2009 of "the U.S. President" doesn't have the same referent as an utterance in 1999 of that same expression, both utterances are assigned their respective referents by the same semantic rule. The circumstances have changed—hence the change in referent—but the semantic rule has stayed the same.

That rule is: if, at time *t*, *x* uniquely has the property of being a U.S. president, then an utterance at *t* of "the U.S. President" refers to *x*.

7.2.1 Incomplete definite descriptions

Some definite descriptions appear to be "incomplete"—that is, they fail to pick out a single object. So, for example, "the bald guy" could pick out any one of many different people. But, like all expressions, definite descriptions are not used in a vacuum; and the context usually supplies the information needed to enrich the mediating concept enough to enable it to single out a single person. So if you and I are in a room and I say "the bald guy is wealthy," it's clear that, so far as I am making a determinate statement, what I mean is that the *contextually salient* bald guy is wealthy.

If phi is a property that obviously has no more than one instance (e.g., the property of being the whole number successor of one or of being the U.S. President), phi's sole instance is contextually salient by default.

7.3 The expressive limitations of indexical-free languages

Indexical-free languages are expressively impoverished; much of what there is to say can't be said in them. The reason is that much information is *perspectival*, and nothing perspectival can be stated in a language that doesn't contain indexicals.

Suppose that I spoke a language that didn't contain any indexicals, but was otherwise just like English. In order to express what I believe concerning the weather in my area, I couldn't say "it is now 60°." I'd have to say "it is 60° in Richmond on Feb. 07, 2009." (And the "is" couldn't be taken as the *present tense* of the verb "to be," since, thus interpreted, it would be an indexical that picked out the time of utterance. The "is" would have to be stripped of any temporal meaning and, thus, be downgraded to an empty grammatical place-holder, like the "it" in "it's raining.") But that utterance wouldn't necessarily express the belief I wanted to express. My believing that *it is 60° in Richmond on Feb. 07, 2009*, is different from my believing that *it is now 60° here*, even though I am in fact in Richmond at the time in question. I could have the one thought and not have the other. I could, after all, not know that I was in Richmond. And my believing either of those things is different from my thinking *it is now 60° here.* One could have any given one of those beliefs without having any of the others. Further, one could *rationally* have any given one of those beliefs without having any of the others. A person who is in Richmond at the time in question can

rationally believe that *it is now 60° here* while believing it false that *it is 60° in Richmond on Feb. 07, 2009.* For the data at one's disposal may warrant the first judgment, but not the second.

So *perspectival* beliefs cannot be expressed without using an indexical. By a "perspictival" belief, I mean one whose truth depends, not only on the characteristics of the objects spoken about, but also on what coordinates those objects have relative to the speaker's framework. So an example of such a thought would be the one I express when, while pointing at some man in a van:

> (a) "Ever since I moved *here*, *that* guy has been stalking me, even though *now* it is clear to me how to deal with the situation, even though *yesterday* it was not."

Supposing that Smith is the man in the van and that I utter those words on June 11, 1999, in Los Angeles, the belief expressed by:

> (b) "Ever since Kuczynski moved to Los Angeles, California, Smith has been stalking Kuczynski for two years; but on June 11, 1999, Kuczynski knows how to deal with that situation, even though he didn't know how to deal with it on June 10."

The belief expressed by an utterance of (b) is true so the objects referred to have the properties attributed to them. But even if those very objects had those properties, an utterance of (a) might be wrong: for the speaker might be pointing at Jones, not Smith, and he might be in Richmond, Virginia, not Los Angeles, California, and so on. So, for the belief expressed by an utterance of (a) to be correct, it is necessary, not only that the objects referred to have the properties attributed to them, but also that those objects occupy certain positions within a frame of reference whose center is speaker himself. A language in which perspectival thoughts couldn't be expressed would have no practical value. For the individual's needs are, almost by definition, to be understood in terms of a frame of reference of which he is center. It might be thought that such a language *could* potentially have great theoretical value. But depending on how this thesis is interpreted, it's either false or trivial. The truths that a theoretician (operating in his capacity as a theoretician) aspires to

identify are indeed non-perspectival, and such truths are expressed by sentences constructed out of expressions (e.g. "Alpha Centauri", "kiwi", "Earth", "electron") whose occurrences are *not* to be understood in terms of this or that person's frame of reference and are instead to be understood impersonally. But there is no way to endow such terms with meaning *except* through the use of terms (e.g. "here", "there", "now") whose occurrences *must* be understood in terms of specific frames of reference. Sentences like "*this* is planet Earth", "*that* is Alpha Centauri", "the animal we are *now* observing is a kiwi" are needed to give meanings to the terms "Earth", "kiwi", and "Alpha Centauri." Of course, such terms can, and usually are, defined by describing, as opposed to directly indicating, their referents. One can define "Alpha Centauri" by describing it as "the star nearest the Earth, apart from the Sun", and one define "Venus" as "the planet closet to the Sun, after Mercury." But such descriptive definitions would be impossible were it not for non-descriptive, purely demonstrative definitions, such as "*this* planet is the Earth" and "*that* thing is the Sun."

Of course, one could imagine a language L that contained no perspectival terms (or "indexicals" as they're typically called) and whose expressions in terms of perspectival terms belonging to some *other* language L*. But this only proves our points, since it is only thanks to perspectival sentences belonging to some language *other* than L that L even exists to give those who speak it the means to state impersonal truths. So the relevant body of semantic rules, and thus, in effect, the relevant language, consists of those belonging to L *as well as* those belonging to L*.

Also, theoretical truths are learned on the basis of the information embodied in one's sense-perceptions and post-perceptual beliefs; and that information is necessarily perspectival. So L would be completely useless *except* as a way of stating discoveries already made. The sentence-utterances involved in the process of making such a discovery would have to belong to a language, such as L*, that contained indexicals. Thus, the limitations of indexical-free languages are consequences of fundamental epistemological truths; and the study of indexicals needn't be the sterile, zero-sum enterprise that many investigations in semantics are generally believed to be.[155]

7.4 Tokens, types, and context-sensitivity

No sentence-*type* containing a context-sensitive component is either true or false; for no such sentence-type says anything. It is *tokens* of "I am now tired" that make statements; the corresponding sentence-type does not do so. What a token of "I am tired now" affirms depends on facts about the context of utterance. If Smith's the one who's speaking, and it's 3:00 P.M., such an utterance is true just in case Smith is tired at 3:00 P.M. If Jones is the speaker and it's 4:00 P.M., such an utterance is true just in case Jones is tired at 4:00 P.M. Thus the semantic rule that assigns a proposition to such a token does so on the basis of the facts about the context of utterance. And the rule in question is clearly this: "If, at time *t*, person *p* tokens the sentence-type "I am tired," then the proposition thereby affirmed is true exactly if *p* is tired at *t*." The same thing *mutatis mutandis* is true of all context-sensitive expressions. So the rule that assigns a proposition to an utterance of "that man is married to that woman" is: if, in the context of utterance, *x* is a uniquely salient man and *y* is a uniquely salient woman, a token of the sentence-type "that man is married to that woman" affirms a proposition that is true if and only if *x* is married to *y*.

So context-sensitive sentence-types aren't true or false; they don't, in and of themselves, bear propositions. But they have an important semantic role: the identity of the proposition meant by a token of such a type depends on the identity of the type. A token of "I am tired" means one thing; a token of "you are tired," uttered by the same person at the same time, means something else; and that difference obviously stems from the fact that, because different sentence-types were tokened, the semantic rule that assigns a proposition to the one token is different from the rule that assigns a proposition to the other.

This can all be distilled into the following principle: where context-sensitive sentences are concerned, the meaning of a token is a proposition, and the meaning of type is a *rule* that assigns a proposition to one of its tokens on the basis of facts about the context in which that token occurred.

Long story short: the meanings of sentence-tokens are propositions and the meanings of sentence-types are rules that assign propositions to their tokens, usually on the basis of facts about the context of tokening.

7.5 Ambiguity and context-sensitivity (revisited) and the type/token distinction (revisited)

Consider the sentence-type "that person is a professor." Is that sentence true or false? No. Some tokens of it are true and some are false. This is because what a given token of that sentence says is a function of the circumstances. If I utter it while pointing at Bob, I'm attributing the property of being a professor to Bob. If I utter it while pointing to Sally, I'm attributing that property to Sally, not Bob. So what it is that I'm affirming in the one case is different from what it is that I'm affirming in the other case. But that isn't because the sentence "that person is a professor" is ambiguous. That sentence is not ambiguous: it has just one meaning. But that meaning is not a proposition; it isn't something that is true or false. That meaning is a rule. That rule in its turn assigns meanings to tokens of that sentence. Those meanings are true or false; those meanings are propositions.

Remember that expression-types are what result when properties *per se*, as opposed to their instances, are assigned meanings. Of course, the rules that assign meanings to properties are semantic rules, since anything that assigns meaning to anything is *ipso facto* a semantic rule. So the sentence-*type* of which the following ink deposit—"that guy is a professor"—is a token is what results when some semantic rule assigns a meaning to some property. Let R_1 be the semantic rule in question. Let R_2 be the meaning that R_1 assigns to the just-mentioned property. R_2 is *itself* a semantic rule. But, whereas R_1 assigns a meaning (a rule) to some property, R_2 assigns meanings to instances of that property. So R_2 assigns meanings to particular instances of the morphology had by the following ink deposit ("that guy is a professor"). The meanings that R_2 assigns to those tokens are not themselves rules; they are propositions—they are things that are true or false.

There are thus very different sorts of semantic rules. There are those that assign meanings to properties, and there are those that assign meanings to instances of those properties. The meaning that was assigned to the property we were just discussing is itself a semantic rule. As we'll now see, *every* meaning that is assigned to a property (as opposed to a property-instance) is itself a semantic rule. Consider the following ink deposit: "Barack Obama." There is a semantic rule that assigns a meaning to the *property* of having a morphology similar to that ink deposit. The meaning assigned to that property is *itself* a

semantic rule. *That* rule assigns a meaning to each instance of the property in question. It assigns Barack Obama, the person, to any such instance. Thus, any such instance is an expression that picks out Barack Obama.

In the previous section, we said that semantic rule is something which assigns a meaning to a property. That's correct, but incomplete. The right definition is this: a semantic rule is something which assigns a rule to a property that in its turn assigns meanings to instances of that property. Let us now move onto slightly less abstruse material.

7.6 Do any expressions have one-dimensional semantics?

The question arises: What about sentence-types that don't contain context-sensitive components? Where they are concerned, does token-meaning coincide with type-meaning?

First of all, in natural languages, there are no such sentence-types. Every sentence contains a tense-marker. And in virtue of containing a tense-marker, a sentence is such that the proposition expressed by any given one of its tokens is a function of (*inter alia*) when that token occurs. If you say "the U.S. economy is fairly stable," whether you are speaking the truth or not depends on when you say it.

There are some *apparent* counterexamples to this. For all intents and purposes, any two tokens of:

> (IA)[156] "The interior angles of a Euclidean triangle add up to 180°"

express the same proposition. But, from a strictly semantic perspective, IA *is* context-sensitive, and non-simultaneous utterances of it *don't* encode the same proposition. If it were believed that mathematical reality were as volatile as the stock-market, we would without hesitation regard IA as being in the same category as patently context-sensitive sentences such as:

> (BH) [157] "Bill's holdings add up to $180,000,000."

And we'd have no more temptation to regard the tense-marker in IA as inert than we'd have to regard its counterpart in BH as inert. What this shows is that, to the extent that the tense-marker in tokens of IA are doing nothing, it's only because, our beliefs about mathematics being what they are, we choose to see such utterances as expressing atemporal propositions.

Nonetheless, many semanticists hold that at the level of literal meaning IA is context-insensitive. In other words, they hold that IA is what Quine (himself such a semanticist) refers to as an "eternal sentence." (*S* is an eternal sentence if there is some one proposition *P* such that any two of *S*'s tokens express *P*.) If only for argument's sake, let's suppose that there are eternal

sentences.

It's tempting to hold that an eternal sentence's meaning is identical with those of its tokens. After all, the distinction between type-meaning and token-meaning seems quite hollow *except* where context-sensitive expressions are concerned.

But this reasoning is spurious. The tense-marker on the occurrence of "add" in (IA) has the same semantics that it does in BH, where the tense-marker is obviously doing real work. That's why utterances of BH are true on Monday (before the market crashed) and false on Wednesday (after the market crashed). Therefore, different tokens of BH express different propositions. Therefore, the meaning of BH—the sentence-type—isn't some proposition, and is instead the rule:

(BHSR) If *t* is a token of BH that is uttered at time T, *t* is true iff, at *T*, Bill's various holdings add up to $180,000,000.

But, of course, for any time *t*, the meaning of a token of BH that is produced at *t* is a proposition that is true just in case.

(BHT) At *t*, Bill's various holdings add up to $180,000,000.

Given that the occurrence of "add" in any given token of IA has the same literal meaning as its counterpart in any given token of BHT, it follows that, from a narrowly semantic perspective, IA is quite as context-sensitive as BHT. To be sure, there is obviously a sense in which IA is, whereas BHT is not, context-*in*sensitive. But IA's context-insensitivity, we must conclude, is a thoroughly pragmatics-based affair.

Obvious extensions of this reasoning show that *all* expression-types have *rules* that assign meanings (or referents) to their tokens and, therefore, where any expression is concerned, type-meaning diverges from token-meaning. The meaning of the expression-type "that man" is: a token of "that man," uttered in context *C*, refers to *x* if, in *C*, *x* is uniquely a salient man. If, in a given context, Smith is such a man, then a token in that context of "that man" refers to Smith; if instead Jones is such a man, it refers to Jones; etc. The same is true of "the current President." In one context (the year 1992), tokens of that expression refer to Bill Clinton; in a different context (the year 2009) it refers to Barack Obama.

Some would be tempted to say that, because they have fixed referents, proper names (e.g., “Bill Clinton”) don’t have a two-dimensional semantic structure. This would be a mistake. Expression-tokens are physical entities—bursts of noise, deposits of ink, etc. Expression-tokens are therefore perceptible entities and are thus capable of transmitting information. Expression-types, on the other hand, are abstract entities and are thus inherently unsuited to be vehicles of communication. Let us develop these points.

Expression-types are properties. Why are they properties? Tokens of a type are instances of it. Anything of which there are instances is *ipso facto* a property. So expression-types, unlike expression-tokens, are properties and are therefore non-spatiotemporal entities. And, as we just noted, it makes no sense to suppose a non-spatiotemporal entity could mediate information or, therefore, could in any significant sense be a symbol. Also, given the profound metaphysical differences between tokens and types, it would be theoretical arbitrariness of the worst kind to suppose that types could discharge the same semantic functions as their tokens. We must therefore assume that, whereas tokens of “Bill Clinton” refer to Bill Clinton, the corresponding type does not refer to Bill Clinton; and we must also assume that “Bill Clinton,” the expression-type, has for its meaning a (constant) function that assigns Bill Clinton to any given one of its tokens. In general, proper names, no less than indexicals and definite descriptions, have two-dimensional semantics.

The semantic rule corresponding to a proper name is a constant function, whereas the semantic rule corresponding to an indexical or definite description is not a constant function. All utterances of “Bill Clinton” refer to Bill Clinton, but not every utterance of “the current U.S. President,” or of “that guy over there,” so refer. This has encouraged the erroneous view that “Bill Clinton” itself, the expression-type, refers to Bill Clinton.[158]

7.7 Ambiguity and context-sensitivity (re-revisited) and the type-token distinction (re-revisited)

Remember that expression-types are what result when properties *per se*, as opposed to their instances, are assigned meanings. Of course, the rules that assign meanings to properties are semantic rules, since anything that assigns meaning to anything is *ipso facto* a semantic rule. So the sentence-*type* of which the following ink deposit—"that guy is a professor"—is a token is what results when some semantic rule assigns a meaning to some property. Let R_1 be the semantic rule in question. Let R_2 be the meaning that R_1 assigns to the just mentioned property. R_2 is *itself* a semantic rule. But, whereas R_1 assigns a meaning (a rule) to some property, R_2 assigns meanings to instances of that property. So R_2 assigns meanings to particular instances of the morphology had by the ink deposit ("that guy is a professor"). The meanings that R_2 assigns to those tokens are not themselves rules; they are propositions —they are things that are true or false.

There are thus very different sorts of semantic rules. There are those that assign meanings to properties, and there are those that assign meanings to instances of those properties. Consider the following ink deposit: "Barack Obama." There is a semantic rule that assigns a meaning to the *property* of having a morphology similar to that ink deposit. Let *PBO* be that property. The meaning assigned to *PBO* is *itself* a semantic rule. In other words, the meaning of *PBO* is given by the statement that:

B_1: Any given token of *PBO* refers to Barack Obama.

Given some specific token *t* of "Barack Obama," the semantic rule for *t* is:

B_2: *t* refers to Barack Obama.

B_1 and B_2 are very different rules. Unlike B_2, B_1 doesn't say anything about any *specific* token of "Barack Obama" or any other expression. The need for a two-dimensionalist approach to semantics is embedded in the very concept of what an expression is.

8.0 Logical form[159]

Some statements that, given their grammatical forms, appear to be about objects are in fact about properties. We saw this in Chapter 1, and we'll see it again in Chapter 7 when we discuss Frege's ground-breaking insights into arithmetical statements (e.g., "2 + 3 = 5"). Frege was the first to see clearly that logical and grammatical form may diverge—he was the first to grasp the very idea of such a divergence. This insight of his is embodied in his statement that the sentence:

(WM) "Whales are mammals,"

is not about whales. As paradoxical as it may seem, he was right. WM says that *if* an object is a whale, *then* it's a mammal. But there is no specific object *x* such that WM says that *x* is a mammal. *A fortiori* there is no specific whale *x* such that WM says that *x* is a mammal; and for any number *n*, no matter how high, there are no whales $x_1....x_n$ such that WM says that x_i ($1 \leq i \leq n$) is a whale. So just as Frege said, WM isn't about whales.

WM makes a statement, not about whales, but about the property of being a whale. It says that:

(WM*) the property of being a whale has the property of being instantiated only by mammals.

WM* ***perspicuously*** represents WM's meaning. In other words, WM* represents WM's logical form. WM*'s grammatical form diverges from WM's grammatical form. (For a statement to be "perspicuous" is for it to be clear in a *logical*, as opposed to a psychological, sense. The statement "if n is the number of Americans, and m is the number of cars collectively owned by American citizens, then n/m=2.2" makes it clear what is meant by "the average American citizen owns 2.2 cars" and is thus a *perspicuous* affirmation of the proposition meant by the latter. But the latter is probably *psychologically* easier to absorb than the former.)_

The reason that WM's logical and grammatical forms pull apart is that WM contains a ***quantifier.*** (Examples of quantifiers are "all birds," "no man," "most whales," "three birds," and "some individuals." We'll define the term "quantifier" in a moment.) Given *any* statement of English, (or any other

non-artificial language) that contains quantifiers, the logical and grammatical forms of that sentence diverge. A quantifier is an expression having the property that, if a sentence contains it, that sentence is *ipso facto* to the effect that the ***extension*** of one property has a certain degree of overlap, ranging anywhere from no overlap to total overlap, with the extension of some other property. “All birds have beaks” says that the extension of the property of being a bird is a subset of the extension of the property of having a beak. “Only birds have beaks” says (falsely) that the extension of the property of having a beak is a subset of the extension of the property of being a bird.

In virtue of containing a quantifier, a sentence is about properties, not specific objects. Of course, a quantified (quantifier-containing) statement can *also* be about individuals. But it isn’t *in virtue of* containing a quantifier that a sentence concerns individuals; and it *is* in virtue of containing a quantifier that a sentence concerns properties.

In natural language, quantified sentences are pseudo-objectual statements. They appear to be about objects, but are really about properties. That is why they must be reparsed if their logical forms are to be exposed. It immediately follows that a statement’s grammatical form may diverge from its logical form. Frege discovered both facts and, therewith, created modern analytic philosophy.

8.1 Contextual definition

The fact that grammatical and logical form sometimes diverge is related to the fact that some expressions are to be defined contextually. To define an expression contextually is to show how, in virtue of containing it, a sentence's meaning is affected. So, for example, "someone" doesn't refer to anyone. It doesn't refer to John or Sally or Jane. That is why, given any proper name *N*, the statement:

(NS) ‹*N* doesn't snore ›,

is ***compatible*** with:[160]

(SS) "Someone snores."

Thus, the meaning of "someone" isn't given by some rule that pairs it off with this or that individual. In other words, there is no individual N such that the semantic rule for "someone" is:

(WRS)[161] ‹Someone has psi› is true just in case N has psi.

Rather, the meaning of "someone" is given by the statement that:

(RS) ‹Someone has psi› is true just in case the property of being a psi is instantiated.[162]

Thus, the logical form SS is clearly displayed by the sentence:

(PS)[163] "The property of being a snorer is instantiated."

PS's logical form therefore coincides with its grammatical form. Since SS has a different grammatical form from PS, it follows that SS's logical form diverges from its grammatical form. The divergence between SS's grammatical and logical forms is obviously a consequence of the fact that "someone" must be defined contextually.

8.2 Contextual definition: Its scope and limits

Nonetheless, it would be an overstatement to say that *whenever* a sentence contains an expression that must be defined contextually, its grammatical form pulls apart from its logical form. As we're about to see, *every* expression is to be defined contextually. Obviously grammatical form doesn't *always* pull apart from logical form. Therefore, it isn't *always* the case that, in virtue of containing an expression that must be defined contextually, a sentence's logical and grammatical forms diverge.

How can it be said that all definitions are contextual definitions? Aren't there also ***denotative definitions?*** (A denotative definition of an expression *E* says what *E* means by saying what it denotes.) Isn't "Socrates" defined denotatively? Yes, it is. There is some object *x* such that one says what "Socrates" means if, and only if, one says that "Socrates" refers to *x*.

But in saying of some object *x* that "Socrates" picks out *x*, one is saying that, in virtue of having the form ‹Socrates has psi›, a sentence *S* is to the effect that *x* has psi. "Socrates" refers to Socrates because, the remaining semantic rules of English being what they are, if you wish to attribute the property of being wise to Socrates, you can do so by saying "Socrates is wise"—because, in general, for any property psi, if you wish to attribute psi to Socrates you can do so by saying ‹Socrates has psi.› If, in saying "Socrates was wise," you were really saying that it was Aristotle, not Socrates, who was wise, then "Socrates" wouldn't refer to Socrates, at least not in that context. In general, to say that *E* refers to *O* is to make a statement about effects that containing *E* has on a given sentence's truth-conditions. More precisely, it is to say that, in virtue of having the form ‹*E* has psi›, a sentence attributes psi to *O*, for any property psi.

To put this last point another way: it is say that, in virtue of having the form ‹…*E*…›, a sentence attributes…*x*…to *E*, where…*x*...may be any property. So "Plato" refers to Plato because "Aristotle admired Plato" attributes to Plato the property of being a thing *x* such that Aristotle admired *x*. If "Aristotle admired Plato" attributed the property of being hated by Mussolini to Gordon Brown, then, supposing the remaining semantic rules of English not to have been changed, "Plato" would (at least in that context) refer to Gordon Brown and "Aristotle" would (same qualification) refer to Mussolini. But the logical forms of sentences of the form ‹E has psi› don't necessarily pull apart from their grammatical forms. Since there is some

object *x* (namely, Socrates) such that "Socrates is tall" says that *x* is tall, and since ‹*x* is tall› has the same form as "Socrates is tall," the fact that "Socrates" is to be defined contextually does *not* entail that logical and grammatical form ever diverge.

So it is only because *certain* expressions are to be defined contextually that such divergences occur. But which expressions? *Those that cannot also be defined denotatively.* Whenever an expression can be defined denotatively, a sentence's logical form will not, by virtue of that sentence's containing that expression, diverge from its grammatical form. (But, of course, that sentence's logical form may diverge from its grammatical form for some other reason. Thus, the logical form of "Socrates saw someone" diverges from its grammatical form; but the reason for this is that it contains the word "someone." The occurrence of "Socrates" in that sentence isn't what induces that divergence.)

Let *E* be an arbitrary expression that can be defined denotatively. In other words, suppose there to be some object *x* such that *E* is defined by saying that it refers to *x*. The logical form of ‹*E* has psi› is: *x has psi.* In general, *to the extent* that the logical form of a sentence containing *E* is determined by its containing that expression, that sentence's logical and grammatical forms coincide. But, since such a sentence's logical form isn't determined *only* by its containing *E*, and is also a function of the semantics of the other expressions occurring in it, its grammatical form may still diverge from its logical form. After all such a sentence may contain an occurrence of "someone" or "nobody" or some other expression that induces such a divergence.

8.3 Scope, propositional attitudes, and the de dicto/de re distinction.

A propositional attitude is an attitude towards a proposition. My wish that Smith would leave is an attitude (one of desire) towards a proposition (*that Smith soon goes away*). The sentence:

(1) "John hopes that the current U.S. President will fix the economy,"

ascribes a certain propositional attitude to John. The attitude: hope. The proposition: that the current U.S. President will fix the economy. So (1) can be thought of as having subject-predicate form, with "John hopes" being the predicate (the expression that denotes the property being attributed to the subject) and the underlined expression being the subject.

"Know," "believe," "hope," "doubt," "suspect," "want," and "wonder" are examples of verbs that denote propositional attitudes. There are many other such verbs. Any verb whose complement (i.e., whose grammatical object) is an expression denoting a proposition falls into this category.

It should be pointed out that, in the sentence:

(2) "John wants to punch Smith,"

the grammatical complement of "wants" *is*, despite appearances, an expression denoting a proposition. For (2) is a truncated form of the sentence:

(2*) "John wants it to be the case that he punch Smith."

Oftentimes the sentential complements of verbs denoting propositional attitudes are compressed into expressions that very much do not appear to denote propositions. For example, in the sentence:

(3) "John wants a boat,"

the underlined part is an abbreviation of an expression denoting a proposition; for (3) is an abbreviated way of saying:

(4) "John wants it to be the case that he has a boat."

8.3.1 The concept of scope

In:

(1) “John hopes that the current U.S. President will fix the economy,”

the underlined expression falls within the ***scope*** of the expression “John hopes.” All this means is that, in (1), the property expressed by “John hopes” is being attributed to the proposition denoted by the underlined expression. (1) says that a certain proposition, viz., *the current U.S. President will fix the economy*, has a certain property, viz., *that of being such that John wants it to be true*.

In general, the scope of an expression *E*, occurring in a sentence *S*, is the largest expression *E** such that, in *S*, the property expressed by *E* is being attributed to the proposition denoted by *E**.

If an expression *E* falls within the scope of an expression *E**, *E* is sometimes said to “govern” *E**

In the sentence:

(5) “It is possible that Jim has three cars,”

the underlined expression falls within the scope of “it is possible that.”

“It is possible” doesn’t denote a propositional attitude. Thus, an expression may fall within the scope of an expression that doesn’t denote a propositional attitude.

8.3.2 De dicto vs. de re

(1) is ambiguous. A story will make this clear. John lives next to the current (2010) President of the United States, namely Barack Obama. But John doesn't know that Barack Obama is the current President. John mistakenly thinks that Barack Obama is in charge of the Federal Reserve, and John hopes that he will fix the currently moribund U.S. economy. At the same time, John believes, for some strange reason, that the American system of government makes it impossible for a president to have any effect on the U.S. economy, and he therefore hasn't the faintest hope that an economic turnaround would have anything to do with the activities of anyone occupying the office of Commander in Chief.

So there is some specific individual *x* such that, although John doesn't know it, *x* is the current President and such that (1) is true if it is taken as a way of saying that John hopes that *x* will fix the economy. But (1) is false if it is taken to mean that John hopes that an instance of the concept *current U.S. President* will fix the economy.

Thus, (1) is true when read *de re* and false when read *de dicto. De re* means "about the individual," and *de dicto* means *about the concept.* (1) is false if taken to concern John's attitude towards the concept *current U.S. President*, but it's correct if taken to concern the individual who happens to instantiate that concept.

8.4 Frege's generalization of the concept of a function[164]

As we'll see in Chapter 7, what made it possible for Frege to revolutionize logic was his insight that grammatical and logical form diverge; and what made the latter insight possible was his generalization of the concept of a mathematical ***function.***

A mathematical function is a rule that assigns no more than one object to each object falling in a given class. (So "+1" can be thought of as expressing a function or rule that assigns 2 to 1, 3 to 2, etc.)

According to Frege, the occurrence of "snores" in:

(PS) "Plato snores,"

is best represented as the open sentence ‹*x* snores›; and that open sentence

is best thought of as expressing a function that assigns the truth-value *true* to each snorer and the truth-value *false* to each thing that doesn't snore. (For brevity's sake, I'll henceforth use the words "truth" and "falsehood" instead of, respectively, "the truth-value *true*" and "the truth-value *false*.")

Here's the idea. A true sentence results if the variable in ‹*x* is even› is replaced with an expression denoting an even number and false if it's replaced with a number that doesn't denote such a number. ("2 is even" is true and "3 is even" is false.) So we can think of ‹*x* is even› as assigning truth to two, four, etc., and falsity to one, three, etc. For similar reasons, we can see ‹*x* snores› as assigning truth to snorer Bob and falsehood to non-snorer Wilma, and so on.

Frege sees the occurrence of "is taller than" in:

(BTM) "Bill is taller than Mary,"

as being identical with the open sentence ‹*x* is taller than *y*›, and he sees that open sentence as expressing a function that assigns truth-values to *ordered pairs* of objects. So ‹*x* is taller than *y*› assigns truth to the ordered pair <*x*, *y*> if *x* is taller than *y* and otherwise assigns falsehood to that pair.

Frege treats:

(BJMP) "Bill is standing between Mary and John,"

as comprising the open-sentence ‹*x* is standing between *y* and *z*›; and he sees that open sentence as expressing a function that assigns truth to ordered triples of objects—as assigning that truth value to <Bill, Mary, John> just in case Bill is standing between Mary and John.

In this way, Frege was able to assign a *single* form to all ***atomic*** sentences. An atomic sentence is one that, unlike "Bill is tall and Sally is smart," doesn't consist of other sentences and that, unlike "someone snores," doesn't contain quantifiers. Non-atomic sentences are ***molecular.*** A molecular sentence is one that either consists of other sentences or contains a quantifier. Frege was able to show that, ultimately, all atomic sentences have the form ‹*O* has psi. › Contrary to first appearances, BTM has the form: *the ordered pair <Bill, Mary> has psi,* where psi is the property had by such a pair just in case its first member is taller than its second.

Pre-Fregean logicians saw each of PS, BTM, and BJMP as having a

different form from each of the other two, and this made it impossible for them to do anything meaningful in the way of formalizing inferences involving atomic sentences. But Frege didn’t have this problem, since he, unlike them, wasn’t made blind by grammatical surface structure to the underlying structural similarities.

Chapter 4 What Are Propositions?

1.0 The purpose of this chapter

In this chapter, we'll provide abbreviated answers to the questions "what are propositions?" and "what is it for propositions to be true?"[165] Here's what we'll see. Propositions are sets of properties; and a proposition is true just in case the properties that are members of it are instantiated -- so what it is for a proposition to be true is for the members of the set with which it is identical to be jointly instantiated. Roughly, truth is instantiatedness. Less roughly, truth is identical with the property of being a set (of a certain kind) all of whose members are instantiated. Aside from a few brief remarks, I will leave it open whether *all* sets of properties qualify as propositions (hence the parenthetical hedge in the last sentence). But a case will be made that *only* sets of properties are propositions.

1.0 The Argument

Consider the proposition:

1. Tommy is smoking.

Intuitively, this proposition seems to consist of at least three components: first, Tommy (or, in any case, something *individuative* of Tommy, that is, a property P such that anything having P is *ipso facto* identical with Tommy and Tommy alone; second, the property of smoking; and, third, some kind of "synthesis" of the two – some complex constituent that involves Tommy's smoking. (We will try to state precisely what this synthesis involves.)

If 1 is true, then:

A. The property of being identical with Tommy,

is instantiated, and so is:

B. The property of smoking.

But, of course, those two properties could be instantiated in a world where Tommy is not smoking – in a world where, for example, Larry is smoking and a non-smoking Tommy is playing golf. We thus need a third property – a property such that, if a set *S* contains that property along with A and B, then 1 will be true just in case all of *S*'s members are instantiated. That third property is not hard to identify. It is:

C. The property of being identical with something that is smoking and is identical with Tommy.

Before moving on, let's look at the apparatus that we've been developing. Once again, let *S* be a set that contains all and only A, B, and C[166]; and remember that we have identified *S* with 1 (the proposition that Tommy is smoking).

First of all, if C is instantiated, so are B and A. Further, if C, and therefore B and A, are instantiated, then, and only then, is 1 true. It follows that 1 is true if, and only if, each of A, B, and C is instantiated. This is consistent with,

and even corroborates, the thesis that 1 *is S* that and that 1's being true *is S*'s members' being jointly instantiated.

Of course, that does not by itself show that 1 can be identified with *S* or that 1's being true can be identified with *S*'s members' being jointly instantiated. After all, there are infinitely many distinct sets of properties such that 1 is true just in case their members are instantiated. For example, let *Z* be the set containing the following properties and nothing besides: the property of being identical with a smoking Tommy; the property of being an even number; the property of being three sided *if* a triangle. The second two properties are instantiated in every possible world, given that even numbers necessarily exist and given that triangles cannot fail to have three sides. But surely *Z* cannot be identified with 1, the reason being that *Z*'s membership cannot, at least not in a sufficiently natural manner, be aligned with what we know about 1's constituency.

But *S*'s membership *does* so align, as a brief look back makes clear. *S*'s members are (or at least include) A, B, and C. 1's constituents are (or at least include):

(i) Something individuative of Tommy,

(ii) The property of smoking, and

(iii) Something that combines those two things.

Obviously A is individuative of Tommy. (Only Tommy can have that property; and, what is more, he *must* have it.) So far, so good. (ii) and B are identical. So far, still so good.

What about C? That property is instantiated exactly if A and B are both instantiated. So, in a very clear sense of the term 'combine,' that property combines the property of smoking with (something individuative of) Tommy. (In the relevant sense of the term 'combine,' one property 'combines' *n* other properties exactly if the first is instantiated exactly if the other n properties are jointly instantiated.) And we know that, in some way or other, 1 combines (something individuative of) Tommy with the property of smoking. In at least one respect, then, C's relationship to A and B is suggestively similar to 1's relationship to its parts. Moreover, C 'dominates' each of A and B in a way suggestively similar to the way in which 1 dominates its constituents. A's

being instantiated is necessary but not sufficient for C's being instantiated, the same being true of B's being instantiated. Tommy's existing is necessary but not sufficient for 1's being true, the same being true of somebody's being a smoker.

While none of this demonstrates definitively that 1 is identical with *S* or that 1's being true is identical with *S*'s being instantiated, we have, I believe, found enough of a match between 1/1's being true and *S*/*S*'s members' being instantiated to warrant looking into the idea that they might be one and the same.

3.0 The unity of the proposition

How does 1 combine Tommy with the property of smoking? According to Frege, Tommy (or some concept thereof) *saturates* that property (or some concept thereof). (Henceforth, I'll omit the qualification 'or some concept thereof.') But this doesn't help at all. For, as Davidson (1967) observed, the term "saturate" is just another label for the operation in question. So while the Fregean statement:

> (FR) "1 is the result of Tommy's saturating the concept of smoking,"

may well be correct, we don't know what it means.

We are in a position to give the meaning of FR and to say what this mysterious 'saturating' amounts to. We've argued that 1 is a set whose members include the property of being Tommy, the property of smoking, and the property of being a smoking Tommy. 1 is the result of Tommy's 'saturating' the property/concept of smoking in the sense that 1 is identical with a set *S* that consists of those three properties.

For any proposition *X*, to say that *X* is the result of *Y*'s 'saturating' *Z* (for some *Y* and *Z*) is to say that *X* is a set whose members are the property of being identical with *X*, and *Z* itself, and also the property of being identical with something that is identical with *X* and also has *Z*.

4.0 More on the decomposition of propositions

It is indisputable that 1 is true just in case C is instantiated. So why not just identify 1 with C, and 1's being true with C's being instantiated? Why go through the rigmarole that we went through? Why bother with A and B and, in general, any other properties other than C? Because if we identify 1 with C, and 1's being true with C's being instantiated, our theory won't accommodate the very reasonable presumption that 1 has discrete parts corresponding to Tommy and to the property of smoking.

Consider an instance of the property of being a smoking Tommy. Obviously that instance will *involve* Tommy smoking. But it will not, at least in any obvious way, contain a discrete part corresponding to Tommy or a discrete part corresponding to smoking. And what is true of an instance of that property is, quite possibly, true of the property itself. It cannot be taken for granted that C has a *discrete* part corresponding to Tommy or to smoking or to anything else. In fact, being a non-spatiotemporal entity, C won't have discrete parts in any straightforward sense at all.

The impression that C is part-less, or at least lacking in the right kind of parts, is reinforced when we look at what an instance of C is like. Such an instance is, to adapt Kenneth Taylor's (1998) apt expression, a dent or ripple in the quantum; it is a shift in the movement of mass-energy. Of course, that mass-energy is (by hypothesis) sufficient for the existence of Tommy and of smoking and, indeed, of a smoking Tommy. But you would have a devil of a time isolating some part of that distribution that corresponded to Tommy, then distinguishing that from some other part corresponding to smoking, and then finding a third that fused those two, otherwise discrete, entities into a discrete instance of a smoking Tommy.

By contrast, the proposition:

1. Tommy is smoking,

is, of its very nature, neatly parsed into Tommy, the property of smoking, and a third constituent that synthesizes the two just mentioned. The neat articulations found in 1 are found, in some form, in *S*. Those articulations, though found in *S*'s membership as a whole, are not found in C taken by itself, making *S*, but not C, a suitable mirror for the decompositional structure of 1.

States of affairs (or "facts," as they are sometimes called) do not, as Ian Hacking (1975) long ago put it, have the same kinds of “articulations” as propositions. The quantum-ripples on which the truth of a proposition supervenes don’t have anything like the structure of that proposition. Whereas the proposition *Tommy is smoking* clearly has a unique, and relatively simple decomposition into Tommy (or some property/concept individuative thereof), the property of smoking, etc., the state of affairs in which Tommy’s smoking consists – the confluence of psychological, physiological, and, ultimately, sub-atomic disturbances – obviously doesn’t uniquely decompose into Tommy and the property of smoking. In fact, it seems a stretch to suppose that it even has *one* such decomposition.

4.1 A corollary

Consider the proposition:

2. Tommy is snoring.

Let D be the property of being a snoring Tommy. It is obvious that:

1. Tommy is smoking,

and:

2. Tommy is snoring,

have much in common. They have in common the constituent *Tommy* (or, in any case, something corresponding thereto). They have in common the functions (or, if you prefer, properties) *the property of being Tommy who is doing something* and *the property of being a something that is doing something.*

In light of this, consider some instance of D. In other words, consider the kind of property-instance, the kind of mass-energy distribution, on which the truth of 2 supervenes. That instance will, of course, suffice for the existence of Tommy, as will an instance of C. But will those instances, in virtue of that fact, have some *discrete* part in common? Unlikely. In neither case do we have Tommy *simpliciter*. We have *Tommy smoking (while sweating profusely and worrying about his algebra test and his nascent nicotine addiction, all the while talking with his cousin Albert*); in the other we have *Tommy snoring (while having terrible dreams about his future as a second string pitcher on a mediocre team).* There is no such thing, in either case, as Tommy *simpliciter.* There is only Tommy doing this or that, having this or that property. A corollary is that two situations that comprise Tommy do not, in virtue of that fact, have in common some discrete, isolable entity. So while it is true that C and D both comprise Tommy, the one does not, in comprising him, have a discrete constituent in common with the other.

But now we must consider what we've said about the composition of 1 and also what follows about 2, by obvious extensions of what we've said about 1.1 is a set *S* whose members are:

a: the property of being Tommy
b: the property of smoking
C: the property of being a smoking Tommy

By similar reasoning, 2 is a set *S** whose members are:

a: the property of being Tommy,
b*: the property of snoring, and
D: the property of being a snoring Tommy.

S and *S** *do* have well-defined, discrete parts in common. Indeed, they have in common precisely what is had in common by:

1. Tommy is smoking,

and:

2. Tommy is snoring.

If we were to be really precise about it, we'd say that each of *S* and *S** comprised, not only a-c, but also:

d. the property of being a Tommy that does something (i.e., the property being a Tommy that does something or, in any case, has some property),

and also:

e. the property of being a something that does something (i.e., the property being a something that does something or, in any case, has some property).

Let's look at 1 for a second. It immediately and formally implies:

i Tommy smokes;
ii something smokes;

iii Tommy does something;
iv Tommy exists;
v something does something.

And 2 immediately implies:

i* Tommy snores;
ii* something snores;
iii Tommy does something;
iv Tommy exists;
v something does something.

Now let's look at *S* and at *S**. The members of *S* are:

I The property of being Tommy.
II The property of snoring.
III The property of being a Tommy that snores.
IV The property of being a Tommy that does something.
V The property of being a something that does something.

The members of *S** are:

I The property of being Tommy.
II* The property of smoking.
III* The property of being a Tommy that smokes.
IV The property of being a Tommy that does something.
V The property of being a something that does something.

Having identified 1 with *S* and 2 with *S**, everything we know about those propositions – about what they entail, about what their components are, and about how their higher-level components are "forged" out of their lower level components – is accounted for.

5.0 Another desideratum satisfied by our analysis

Many semanticists today hold that Plato and Socrates are *veritable* objects are constituents of propositions that concern them. So according to Kaplan (1989), who is following and agreeing with Russell (1903), Tommy – the person, not some Fregean concept thereof – is a constituent of both 1 and 2. Other adherents of this view are Scott Soames (2003), Nathan Salmon (2005), and Robert Moore (1995). (Opponents are J. Searle (1983), M. Dummett (1973), and – with some very heavy reservations to be stated later – the present author.)

There is a rather serious problem with this view. Consider, not 1 or 2, but rather:

3. Plato snores.

At this point in time, Plato doesn't exist. So if he's a (necessary) constituent of 3 – if, in other words, 3 exists only if Plato does -- then 3 doesn't exist. Salmon and Soames, both of them hard-line direct reference-theorists, deal with this by taking a desperate measure, viz., by saying that 3 exists *and* that its existence constitutively depends on Plato's existence *and* that these two positions can be reconciled with each other --- that, more specifically, they can be reconciled by saying that 3 exists, albeit as a non-existent proposition.

This is not necessary or plausible or helpful. 3 exists. Plato does not. But the property of being identical with Plato, whatever it is, does exist. What is it to be Plato? To be Plato is, perhaps, to be a causal sequence that is initiated by the fertilizing of a certain egg by a certain spermatozoa. To be Plato is, perhaps, to have a soul or mind of a certain kind. What we know is that, if something is Plato, it's being Plato supervenes on something probably having to do with the mass-energy distribution in the cosmos – on its mode of origination, on the quotient of spiritual energy in the particle-interactions that constitute it, on who knows what. In any case, even though Plato does not exist, the property of being identical with Plato does. That's why I can meaningfully, not to mention correctly, say "I am not Plato" or, equivalently, "I don't have the property of being identical with Plato." And it's also why we can meaningfully, and presumably correctly, say that Plato does not exist. For, in making such a statement, we aren't picking out a non-object---which

would, having been picked out, *ipso facto* cease to be a non-object – and then saying of it that it doesn't exist. We are saying of an existing property, that of being identical with Plato, that it is uninstantiated.

But our analysis is still consistent with the heart of direct-reference-theory, this being the idea that, for some *x* such that *x* is identical with Plato, the proposition meant by "Plato snores" is true just in case *x* snores, it being completely irrelevant what properties (other than being a snorer) *x* has. (The Fregean indirect-referentialist holds that for "Plato snores" to be correct it is necessary that, for example, a unique great philosopher to write *The Parmenides* snore.) It is clear why our analysis is consistent on this matter with the viewpoint of the direct referentialist. According to our analysis, 3 is true exactly if the following properties are instantiated:

a. The property of being identical with Plato.
b. The property of snoring.
c. The property of being a Plato who snores.
d. The property of being something that snores.
e. The property of being something that does something (or has some property).

It is readily seen that, on our analysis, there is some *x*, such that *x* is Plato and such that 3 is true exactly if *x* snores, it being irrelevant what other properties *x* has. This is exactly what the direct-referentialist holds. At the same time, our analysis guarantees the existence of the proposition *that Plato snores* -- after all, the *property* of being Plato exists, even though it is now uninstantiated. So our analysis of *Plato snores* gives the direct-reference-theorist the truth-conditions that she wants and that (given certain compelling arguments due to Kripke, 1980; and Soames, 2003, 2004) she probably deserves. At the same time, our analysis doesn't put the direct-reference-theorist in the awkward, logic-bending position of having to say that the proposition that Plato snores exists despite the failure of existence of entities on whose existence its own existence is constitutively dependent.

Chapter 5 Definite Dscriptions and the Nture of Rference

1.0 An old puzzle

Benjamin Franklin (BF) invented bifocals and he was also the first postmaster general. The semantic rules of English being what they are, it follows that:

> (CB) "The inventor of bifocals is identical with the first postmaster general,"

is correct and also that:

> (TB) "The inventor of bifocals is identical with the inventor of bifocals,"

is correct.

But whereas TB is trivial, CB is non-trivial.[167]

The non-triviality of CB is hard to explain if definite descriptions are supposed to be referring terms. As we saw in Section 1.2, the semantic rules of English being what they are, “the inventor of bifocals” refers to BF exactly if ‹the inventor of bifocals has phi› attributes phi to BF. So supposing that “the inventor of bifocals” does in fact refer to BF, it follows that CB attributes the property of being identical with the first postmaster general to BF. And supposing that “the first postmaster general” refers to BF, it follows that CB attributes the property of being identical with the inventor of bifocals to BF. And given both suppositions, it follows that CB says of BF that it has the property of being identical with BF. Which is trivial.

Right now we'll consider Russell's solution to this problem, and we'll then consider Frege's solution. We're considering Russell's solution first because Frege's solution is more easily understood in terms of Russell's than *vice versa*.

2.0 Russell's solution: The Theory of Descriptions

According to Russell's Theory of Descriptions (TD) ‹the inventor of bifocals has phi› says that something uniquely invented bifocals and, moreover, that any such thing has phi. And, according to TD, ‹ the first postmaster general has phi› says that something was a unique first postmaster general and, moreover, that any such thing has phi. It follows that, if TD is right, CB says the same thing as:

> (CB_3) "Something uniquely invented bifocals, and any such thing is a unique first postmaster general; and something was a unique first postmaster general, and any such thing is a unique inventor of bifocals."

It is readily seen that CB_3 is more succinctly expressed as follows (the "R" stands for "Russell"):

> (CBR) "There is some one thing that uniquely invented bifocals and was also a unique first postmaster general."

To sum up, if TD is right, then CB means the same thing as CBR. It is easily verified that, indeed, CB is true exactly if CBR is true. TD thus *seems* to get the right result. (Why "seems"? See Section 11.1) Given that CBR is non-trivial—given that historical knowledge is needed to know it—TD readily explains why CB is non-trivial.

3.0 An aside concerning the difference sentence-type logic and sentence-token logic

On a first reading, go directly to Section 4.0; skip this section.

Why do I say that TD "seems" to get the right result, and not that it *does* get the right result? CB is a sentence. Given only that CB is equivalent with CBR, it doesn't follow that the proposition meant by the one is equivalent with the proposition meant by the other. Strawson (1950) made this clear. (Strawson's insight was popularized by Kaplan (1989). Suppose for argument's sake that the semantic rule for "that person over there is" is: if, in a context *C*, somebody *O* is uniquely a salient person, then an utterance in *C* of ‹that person over there is has phi› has for its meaning some proposition equivalent (if not identical) with *O has phi*; and if, in *C*, there is no uniquely salient person, then such an utterance is neither true nor false (the reason being that the demonstrative doesn't have a referent). But, given that same supposition, the *sentence* ‹that person over there is has phi› would be true exactly if, in the relevant context, somebody *x* were uniquely a contextually salient person and, moreover, any such person had phi.

Thus, given only that Russell's theory correctly models the logic of CB and other *sentences* containing definite descriptions, it doesn't follow that it correctly identifies the *propositions* meant by such sentences. My own view is that, indeed, TD is false. But I also believe, partly for the reasons given here, that Frege's view is false. My alternative to their analysis is put forth in my (2007) book *Conceptual Atomism and the Computational Theory of mind.*

4.0 Frege's solution

Here is Frege's solution to this problem[168][H6] :

> (FR)[169] "The inventor of bifocals" is a referring term. In other words, it picks something out. (Not all expressions do so. [170])That expression refers to a given thing if and only if that thing has the property of being a unique bifocal-inventor. If a given thing didn't invent bifocals then "the inventors of bifocals" doesn't refer to it. If lots of different individuals invented bifocals, then "the inventors of bifocals" doesn't refer to anything. (If 10 people collectively wrote *War and Peace*, it would be wrong to speak of *the* author of that book.) And if a given thing *did* uniquely invent bifocals, then "the inventors of bifocals" *does* refer to it.
>
> So there are *two* dimensions to the semantics of "the inventors of bifocals." There is what it refers to: this is the referent of "the inventors of bifocals." (Sometimes "reference" is used instead of "referent.") And there's the property that a thing must have to *be* its referent; this property is the *sense* of that expression. The referent is BF, and the sense is the property of being a unique bifocal inventor.
>
> Analogues of these points hold with respect to "the first postmaster general." There is its referent (which is BF); and there is its *sense* (which is the property of being a unique first postmaster general).
>
> The reason that TB is non-trivial is that, although the definite descriptions in it have the same referent, they have different senses. A comparison may help. Imagine that you're looking at a certain person in the distance. That person is in fact Ralph, but you don't know it, because he's so far away and also, let us suppose, because he is (for reasons that aren't important) wearing a suit of armor. You see that he is doing jumping jacks. Later Ralph is standing right in front of you and, of course, you see him. He's wearing a plaid shirt. Once again, he's doing jumping jacks.
>
> The first visual perception has the same object as the

second. And each of those perceptions says of that object that he or she is doing jumping jacks. But those two perceptions pick out that object in different ways. In the first case, that object is presented to you as something that has one set of properties (it’s far away, it's wearing a suit of armor, etc.); in the other case it’s presented as having a different set of properties (it’s nearby, it's wearing a plaid shirt, etc.). And that’s why, even though those perceptions have the same object, and even though both attribute the property of doing jumping jacks to that object, those perceptions tell you different things.

Something similar holds of “the inventors of bifocals” and “the first postmaster general.” They have the same object (referent); but they pick that thing out through different bodies of information. That’s why ‹the inventors of bifocals has phi› and ‹the first postmaster general has phi› tell you different things, *even though* they attribute the same property (phi) to the same object (BF).[171]

4.1 Evaluating Frege's solution

This all seems not only plausible, but practically self-evident. But, so far as what Frege says doesn't coincide with TD, what Frege is saying is false and, indeed, incoherent.

Let's consider the two sense perceptions discussed in FR. The one perception tells you that there is something in the distance, that it's wearing a suit of armor, etc., and that, moreover, that thing is doing jumping jacks. So the sense in which Ralph is the *object* of that perception is that your perception *describes* Ralph. Your perception says: "something or other is over there, in the distance, wearing a suit of armor, (etc.)." Who (or what) fits that description? In other words, who has the relevant properties? Ralph. So Ralph is the object of that perception in the sense it *describes* him.

The information borne by that perception has the form "something has such and such properties (it's wearing a suit of armor etc.), and any such thing is doing jumping jacks." Thus, the content of that perception is given by an existence-claim (a claim of the form "something has phi"). Ralph uniquely satisfies that existence-claim. (In other words, he and he alone has the relevant property.) And it is in virtue of his doing so that he is the object of the perception in question.

The same thing *mutatis mutandis* is true of your second perception. The information borne by that perception is along the lines of "something has thus and such properties (it's standing nearby, is wearing a plaid shirt etc.), and any such thing is doing jumping jacks." Ralph is the object of that perception by virtue of the fact that he and he alone has the relevant properties—that he and he alone is a thing standing in the relevant place while wearing a plaid shirt, etc.

In general, when a given thing *x* is picked out *by way of* descriptive information, what's really going on is that some statement of the form "something has such and such properties" is being affirmed (not necessarily verbally), and *x* is what fits the description.

If we bear this point in mind, we find that, so far as it isn't totally incoherent, Frege's analysis collapses into Russell's. Consider the sentence:

(IB)[172] "the inventor of bifocals is tall."

By Frege's own admission, "the inventors of bifocals" picks a given thing

out if and only if that thing fits the relevant description (i.e., if and only if that thing is a unique inventor of bifocals). And if there is a unique inventor of bifocals, that thing, whatever it is, is the thing to which "the inventors of bifocals" refers. So supposing that there is such a thing and that *x* is it, IB describes *x* as being tall.

It goes without saying that, if there is no unique inventor of bifocals, then there is no way that IB can be true. Frege himself emphasizes this last point. (And he's right to do so. There's no way that "the elephant in the corner is eating peanuts" can be true, if there's no elephant in the corner.)

The upshot is that IB is true if and only if:

> (IB*) there is a unique inventor of bifocals and, moreover, that thing, whatever it is, is tall.

But what does TD say about IB? According to TD, the meaning of IB is: *something uniquely invented bifocals, and any such thing is tall.* Which is precisely what FB says. So Frege's theory collapses into Russell's.

By obvious extensions of this, it follows that, if Frege's theory is right, then CB's meaning is identical with that of CBR. So Frege's theory accounts for the non-triviality of CB in the same way as Russell's theory (as TD).

4.2 Evaluating Frege's solution (continued)

For argument's sake, suppose that the prettiest living woman is identical with the most evil living person. Given only that:

> (1) “Smith wants to marry the prettiest woman on the planet,”

is true, it doesn't follow that:

> (2) “Smith wants to marry the most evil person on the planet,”

is true. But why not? Because what (1) really says is that:

> (1*) Smith wants it to be the case that Smith is married to a woman who is prettier than anyone else,

and what (2) really says is that:

> (2*) Smith wants it to be the case that Smith is married to somebody who is more evil than anyone else.

Thus, if (1) is correct, Smith wants a certain proposition to hold, namely, *that Smith is (or will be) married to a woman who is prettier than anyone else*. And if (2) is correct, Smith wants some other proposition to hold, namely, *that Smith is married to somebody who is more evil than anyone else*. Since those two propositions *are* different, the fact that one of them has the property of being desired to be true by Smith, whereas the other lacks that property, does *not* constitute a violation of Leibniz’s Law (*if x=y, x has a given property iff y has that property*). And the “substitution-failure” whereby a truth (namely, (1)), is turned into a falsity (namely, (2)), corresponds to the innocuous fact that, whereas the grammatical complement in each of (1) and (2) of “marry” is a singular term, the grammatical complement is an expression denoting a proposition (as opposed to a person). The occurrence in (1) of “the prettiest woman” is an orthographic expedient —an orthographic stand-in for a more perspicuous but (for that very reason)

unwieldy expression. And surely we shouldn’t let orthographic expedients dictate our semantic beliefs. (Frege himself would be the first to admit this.)

Few authors, if any, deny that apparent exceptions to Leibniz's Law are to be solved in this way. So given that, according to TD, the meanings[173] of (1) and (2) are given by (1*) and (2*), and given that, nuances aside, there's no other way to deal with the datum that (1) and (2) can simultaneously be true, it's odd that authors have continued to countenance Frege's system.

4.3 Special compositionality in relation to Frege's theory

There are other, related problems with that system. According to the principle of special compositionality (SC), what a referring term picks out is a function of what its parts pick out.[174] Bearing this in mind, let's suppose, as before, that the prettiest woman in existence is also the most evil person in existence. And let's also suppose, with Frege, that definite descriptions are referring terms. Given these suppositions, it would seem to follow that the underlined part of:

(3) Smith believes that the prettiest woman in existence is identical with the prettiest woman in existence,

refers to the same proposition as:

(4) Smith believes that the prettiest woman in existence is identical with the most evil person in existence.

But the underlined part of (3) very much seems *not* to co-refer with the corresponding part of (4)[175], that being why, even if the prettiest woman in existence is the most evil person in existence, (3) and (4) can differ in truth-value.

There are only three ways to go here. (i) Deny SC. (ii) Deny that the underlined parts of 3 and 4 refer to different propositions. (iii) Deny that 4 is what results when a referring term in 1 is replaced with a co-referring term.

Understandably, Frege rejects each of (i) and (ii) and thus accepts (iii). (To this extent, his semantics coincides with Russell's, give or take some technicalities relating to the fact that, according to TD, each of 3 and 4 is syntactically ambiguous.) Frege thus says that the definite descriptions occurring in 3 and 4 refer, not to some person, but rather to their "customary senses." What he means is that, in each of 3 and 4, "the prettiest woman in existence" refers, not to Sally or Betty or Ethel, but rather to the property (or concept) of being a woman who is prettier than every other (or some such[176]); and in 4, "the most evil woman in existence" refers to the property (concept) of being a person who is more evil than any other.

An immediate consequence of Frege’s theory is that the word “snow” has three utterly distinct meanings in:

(5) “Snow is white and Jerry believes that snow is white and Fred believes that Jerry believes that snow is white.”

The first occurrence of “snow” in (5) refers to snow. The second refers to the sense of the first occurrence. The third refers to the sense of the second occurrence.

But surely all three occurrences of “snow” in (5) mean the same thing. Frege’s is clearly jumping through hoops to make his bankrupt theory work.

Another immediate consequence of Frege’s theory is that the word “snow” doesn’t refer to snow in:

(6) It is true that snow is white.

But that’s obviously false. (In any case, there's no independent support for this claim.) The occurrence of “snow” in (6) clearly has the same meaning as its counterpart in “snow is white.”

In any case, what does it even mean to say that, in (5), the third occurrence of “snow” *refers* to the sense of the sense of “snow”? If you say “Socrates was bald,” you are attributing baldness to Socrates—not to Aristotle or Plato or anyone else who isn’t Socrates. If you say “Socrates was intelligent,” you are attributing intelligence to Socrates and Socrates alone. Thus, “Socrates” refers to Socrates because, the remaining semantic rules of English being what they are, ‹Socrates has psi› attributes psi to Socrates. By parity of reasoning, if the third occurrence of “snow” in (5) really referred to a second-order sense, then it would make some statement about that sense—it would attribute a property to it. But it doesn’t. It doesn’t say anything about such a sense. Whatever it is that it does in connection with such a sense, it isn’t comparable to what “Socrates was intelligent” does in connection with Socrates.

Finally, to the extent that Frege's views concerning indirect reference are intelligible, they undermine his own semantics. According to Frege, the proposition meant by:

(#) "The prettiest woman in existence is smart,"

is true just in case:

(##) somebody who is smart is uniquely a woman who is prettier than any other.

In general, the proposition meant by:

(*) ‹...the phi... ›,

is true just in case:

(**) something *x* such that...*x*...uniquely has phi.

Let "FR" (short for "Frege-Russell") be the principle ("hypothesis" might be a better word) that the meaning of (*) is given by some proposition that is equivalent (if not identical with) (**). It is to be emphasized that Russell and Frege *both* accept FR.

Here's where they differ. According to Russell's system (TD), ‹the phi› is in the same category as "something," "everyone," etc. According to Frege's system, ‹the phi› is *not* in that category.

Here's the problem. Let *O* be an occurrence of "the prettiest woman in existence" that falls within the scope of four epistemic (or modal) operators. So *O* might be the underlined expression in:

(&) "Smith believes that Sally knows that Larry hopes that Mary is certain that the prettiest woman in existence is smart."

A consequence of TD *and* of Frege's system is that the semantics of the occurrence in (&) "the prettiest woman in existence" is given by FR. But if Frege's system is correct, then that occurrence refers, not to Sally (who we'll assume is the prettiest woman in existence), but to the "customary sense" of the customary sense of the customary sense of the customary sense of the occurrence in (#) of "the prettiest woman.” (Let *O#* be the just-mentioned occurrence.) So, if Frege's system is right, then the semantics of each of *O* and *O#* is given by FR *even though*, whereas the one occurrence refers to Sally, the other refers to a fifth order property (or concept). Thus, a consequence of Frege's theory is that, in respect of their semantics, two expression-occurrences can be functioning in the very same way *even though*

the one occurrence refers to an individual while the other refers to an n^{th} order property (for any whole number *n*). So a consequence of Frege's system is that there can be reference-shifts where there haven't been semantic shifts. Which is obviously incompatible with the principle that, it is in virtue of how a given expression *E* affects the truth-conditions (or meaning) of sentences in which it occurs that *E* refers to this as opposed to that object.[177] We saw in Chapters 3 and 5, and will see in Chapter 6, we saw that for "Socrates" to refer Socrates is for ‹....Socrates...› to be to the effect that ...Socrates...We saw that this condition's being satisfied is necessary and sufficient for that expression's referring to that individual. Thus, for a given expression to have a given referent is for it to have certain effects on the truth-conditions of the sentences in which it occurs.

A corollary is that, where those changes haven't occurred, there hasn't been a shift in reference, a related corollary being that, where they have occurred, there has been such a shift. For any property phi, and for any whole number *n*, the semantics of an occurrence *O^* that falls within the scope of *n* epistemic (or modal or otherwise intensional) operators of the phi is given by FR. Thus, so far as Frege’s clearly contrived and *ad hoc* theory has any intelligible meaning, it’s clearly false.

Chapter 6 Slinging the Slingshot: How a famous argument of Frege's proves his conception of reference to be incoherent

1.0 Is Frege's analysis of reference tenable?

In this chapter we'll find that Frege's analysis of reference---which, on the face of it, is not only reasonable, but downright truistic---is indefensible. Frege's analysis has consequences too spurious for a reasonable person to accept (e.g., that nobody has thoughts about anything, that my typing this sentence is identical with your reading). But rather than rejecting Frege's analysis, philosophers chose to accept these consequences. Why did they do so? First, they couldn't come up with a viable alternative to Frege's analysis. Second, absurd though they are, these consequences supported positions that, independently of their views about semantics, these philosophers had made their life's work to develop and that desperately needed reinforcement. Third, Frege's analysis, preposterous though its consequences be, very much *seems* cogent.

In this section, we'll identify the flaw in Frege's argument, thus stripping its credibility and preventing it from retarding philosophical progress any more than it already has.

2.0 The Slingshot

A "slingshot-style" argument is one to the effect that:

> (i) logically equivalent expressions can't have different referents,
> (ii) intersubstituting co-referring terms preserves truth-value, and
> (iii) given (i) and (ii), it follows that (iii) some class that would seem to comprise several objects (for example, the class of things that are referred to by true sentences) in fact only comprises one object (for example, the property of being true).

In a moment we will present such an argument. That argument is based on the following four sentences
:

> (1) Mozart was a composer.
> (2) The class of all things *x* such that Mozart was a composer and *x* is self-identical is identical with the class of all things *x* such that *x* is self-identical.
> (3) The class of all things *x* such that Bill Clinton was U.S. President in 1997 and *x* is self-identical is identical with the class of all things *x* such that *x* is self-identical.
> (4) Bill Clinton was U.S. President in 1997.

Here is the argument ("SS" is short for "slingshot")

> (SS[178]) (1) is logically equivalent with (2). Therefore, (1) and (2) must co-refer, supposing that either refers to anything. (It is being assumed, very reasonably, that logically equivalent expressions cannot have different referents.)
> (3) is what results when a referring term in (2) is replaced with a co-referring term.) ("The class of all things *x* such that Mozart was a composer and *x* is self-identical" refers to the class of all self-identical objects, and so too does "the class of all things *x* such that Bill Clinton was U.S. President in 1997

and *x* is self-identical.") Since intersubstituting co-referring terms preserves reference, it follows that (2) and (3) co-refer, supposing that either refers to anything.

(3) is equivalent with (4). Therefore, (3) and (4) must co-refer, supposing that either refers to anything.

Therefore, (1) and (4) must co-refer, supposing that they refer to anything.

If the occurrences in (1)-(4) "Mozart was a composer" and "Bill Clinton was U.S. President in 1997" are replaced with occurrences of *any* two true sentences, the argument that results is cogent. Thus, any two true sentences co-refer, supposing that any true sentence refers to anything. (For the same reason *mutatis mutandis*, any two false sentences co-refer, supposing that any false sentence refers to anything.)

The property of being true is the one semantically significant thing that (1) and (4) have in common. So, supposing that sentences refer to anything, (1) and (4) must refer to the property of being true, and true sentences must in general be assumed to refer to the property of being true.[179]

3.0 One problem with SS and with all other slingshot-style arguments

Like all slingshot-style arguments, SS fails to establish its conclusion, the problem with it being that it embodies a spurious conception of reference---one that, were it correct, would entail that reference was indistinguishable from quantification.

Suppose for argument's sake that "the class of all things *x* such that Mozart was a composer and *x* is self-identical" is functioning as a device of reference. In that case, it's picking out the class of self-identical objects, and what (2) is saying about that class is that it's identical with itself; in other words, (2), thus interpreted, says that:

> (2*) The class of self-identical objects is identical with the class of self-identical objects.

But (2*) is not equivalent with (1). Therefore, if the definite descriptions in (1) and (2) are assumed to be referring terms—which is precisely what they *are* being assumed to be—(2) isn't equivalent with (1). For the same reason *mutatis mutandis*, (4) is not equivalent with (3) if "the class of all things *x* such that Bill Clinton was U.S. President in 1997 and *x* is self-identical" is taken to be a device of reference.

In order for (2) to be equivalent with (1), the former must be taken to be to the effect that:

> (2#) There exists a class to which a thing belongs iff Mozart is a composer and that thing is self-identical; and there exists a class to which a thing belongs iff that thing is self-identical; and the first class is identical with the second class.

But if (2) is interpreted in this way, the definite description occurring in it is a quantifier, not a referring term. And in that case (3) is *not* what results when a referring term is replaced with a co-referring term. (It's what results when a non-referring term is replaced with a non-referring term.)

So SS is a failure.

Note: Some would disagree with our contention that, when read attributively, occurrences of definite descriptions are functioning as

quantifiers not as referring terms. Some, in other words, hold that there is such a thing as "attributive" reference. This is the position of Frege and his followers (e.g., Church). In Section 7.0, we'll see why Frege's position is both indefensible and unnecessary.

4.0 Is the conclusion of SS correct? Do true sentences refer to truth-values?

In this section we'll see that, in respect of its semantic properties, “snow is white” bears little resemblance to "the property of being true." And we'll see that, for this reason, it is implausible, if not incoherent, to hold that "snow is white" (or any other true sentence) co-refers with "the property of being true" (or "*the True*" or "the truth-value *True*").

Let’s start with some platitudes about reference. To say that expression R refers to object *O* is to say that, for any property phi, ‹*R* has phi› attributes phi to *O*. If you want to say of Benjamin Franklin (BF) that he’s bald, you can do so by saying “Benjamin Franklin is bald.” (Of course, BF is no longer with us. But to set aside some extremely intricate and, in this context, irrelevant issues relating to tense, let us pretend that Franklin is alive and that we can speak of him in the present tense.) And (the remaining semantic rules of English being what they are) if you *couldn’t* attribute baldness to BF by uttering that sentence, then “Benjamin Franklin” wouldn’t refer to BF. If, in saying “Benjamin Franklin is bald,” one was attributing baldness to *Aristotle*, then, in that context at least, “Benjamin Franklin” would refer to Aristotle.

Let’s generalize these points. For any property phi, “Benjamin Franklin” refers to BF (and not to Aristotle, for example) because ‹Benjamin Franklin has phi› attributes phi to Benjamin Franklin; and if one *couldn’t* attribute phi to Benjamin Franklin by saying ‹Benjamin Franklin has phi›, then “Benjamin Franklin” would not refer to BF. Further, if, in uttering that sentence, one was attributing baldness to Aristotle (but not BF), then “Benjamin Franklin” *would* refer to Aristotle (but not Benjamin Franklin). Thus, for “Benjamin Franklin” to refer to BF it is both *necessary* and *sufficient* that ‹Benjamin Franklin has phi› attribute phi to BF. The fact that “Benjamin Franklin” refers to BF is *identical* with the fact that ‹Benjamin Franklin has phi› attributes phi to Benjamin Franklin, for any property phi.

In general, expression *E* refers to *O* if and *only* if, for any property phi, ‹*E* has phi› attributes phi to *O*. *E*’s referring to *O* may be *identified* with E’s being such that, in virtue of containing it, a sentence *ipso facto* attributes a property to *O*.

Bearing these points in mind, let's revisit the question: Do sentences refer to truth-values? Does "snow is white" refer to the property of being true (which, in this context, we're assuming for argument's sake to be identical

with the truth-value *true*)? No. In virtue of having the form, ‹…the property of being true…›, a sentence is about the property of being true. (“Everything Bill says has the property of being true” says of the property of being true that it is instantiated by everything that Bill says.”) But it isn't the case that, in virtue of having the form ‹…snow is white…›, a sentence is about the property of being true.

Of course, “if snow is white, then snow isn’t green” can, like any other sentence, be interpreted as concerning the property of being true—it can be interpreted as saying that the property of being true is a characteristic of the proposition that snow isn’t green if it’s white. But given how artificial it is to paraphrase "if snow is white, then snow isn't green" in this way, it's clear that, in some respect that is deeply relevant to their semantics, "the property of being true" and "snow is white" aren't comparable. Which, in its turn, suggests that it isn’t *in virtue of* its containing an occurrence of “snow is white” that “if snow is white, then snow isn’t green” concerns that property.

But even if, in virtue of having the form ‹…snow is white…›, a sentence *is* about the property of being true, the *way* in which, in virtue of having that form, such a sentence concerns that property is obviously different from the *way* in which, in virtue of having the form ‹…the property of being true…›, a sentence concerns the property of being true. A story may help make this clear. You are my philosophy professor and I am taking an exam that consists of just one question: I must identify at least one property that all propositions of the form *x is even iff x is divisible by two* have in common. I duly say (or write) "the property of being true." I pass. But I would have failed if I had instead said "snow is white" (or "grass is green"). In any case, you'd have no reason *independently of Frege's argument* for passing me. Which, taken together with the points made in the previous section, suggests that you'd have no reason at all for passing me.

It's hard to find any reason, apart from Frege's argument to the contrary, to hold that "snow is white" refers to the property of being true. Which, given how far Frege's argument to the contrary falls short of being probative, suggests that "snow is white" does not so refer.

4.1 Can sentences coherently be stipulated to have truth-values for their referents?

Some (e.g., Church 1943) have *stipulated* that true (false) sentences refer to the property of being true (false). But some stipulations are incoherent, and this particular stipulation falls into that category. If I stipulate that "Gloxo" is to unambiguously refer to the number two, I cannot coherently stipulate that it is also to refer to the number of sides had by an arbitrary pentagon.

Bearing this in mind, consider the following variant of the just-told story. You are my geography professor, and you've just given me a test containing just one question: identify the state that lies on the Northern border of California. Instead of saying "Oregon," which is the right answer, I say "snow is white." You duly fail me.

Of course, if you and I had agreed that, by our stipulation, "snow is white" is to refer to Oregon, then "snow is white" would have been an acceptable answer. But *to the extent that* "snow is white" would, for that reason, have been an acceptable answer, it obviously wouldn't have been functioning in the same way as the homonymous *sentence* "snow is white." And it is thus incoherent to stipulate that "snow is white" is *unambiguously* both an English sentence (to the effect that a certain crystalline substance has a certain chromatic property) *and* is also an English expression that refers to the state of Oregon. For the same reason *mutatis mutandis*, it is incoherent to stipulate that the English sentence "snow is white" *unambiguously* refers to the property of being true.

4.2 To what, if anything, *do* sentences refer?

Given that "snow is white" doesn't refer to the property of being true, to what, if anything, *does* it refer? It seems to me that, if it refers to anything, that sentence refers to the proposition that snow is white. In virtue of having the form ‹...snow is white... ›, a sentence is *ipso facto* about the proposition that snow is white. “If snow is white, then snow isn’t green” says that the proposition that snow is white entails (or otherwise has as a consequence) the proposition that snow isn’t green. "It is not the case that snow is white" says that the proposition that snow is white has the property of being false. "Grass is green and snow is white" says that the proposition that snow is white has the property of being a thing *x* such that both *x* and the proposition that grass is green are true. So given the reasonable presumption that *E* refers *O* if, in virtue of having the form ‹...*E*... ›, a sentence is to the effect that...*O*..., it seems to follow that "snow is white" refers to the proposition that snow is white.

And even if it doesn't follow, we'd be doing less damage to the concept of reference by saying that "snow is white" referred to the corresponding proposition than we would be if we said that it referred to its truth-value.

A brief detour through the philosophy of science will clarify this last remark. It often happens that, in order for a discipline to move forward, the scope of a given term must be expanded. To be of use to science, terms such as "temperature" and "acid" must be given meanings that, although consistent with the pretheoretic judgments of laypersons, have a much wider sphere of application than do the concepts associated with those terms by the laity. By defining "temperature" as (I'm simplifying) "mean kinetic energy," the jurisdiction of the former term is increased even though its meaning is otherwise unchanged. Such redefinitions are sometimes referred to as ***precisifications*** or ***delineations***.

It's a distinct possibility that, in saying that sentences are referring terms, one is precisifying the term "reference," not identifying its existing meaning. Supposing this were true, it's still wrong (or, in any case, bad in some relevant way or other) to say that sentences refer to their truth-values, and it's *less* wrong to say that they refer to the propositions they express. In order to be a precisification of a given term *T*, as opposed to wholesale redefinition of it, an extension of *T*'s jurisdiction must be consistent (for the most part) with *T*'s existing meaning. (I say "for the most part" since pretheoretic judgments

are sometimes revised in light of their own theoretical off-shoots. There are, or could be, objects *x* such that *x* is cold according to pretheoretic intuition even though, if temperature is identified with mean kinetic energy, *x* is extremely hot.) What we know about the *existing* meaning of "refer" is that (the remaining semantic rules of English being what they are) it is in virtue of the fact that "Socrates" refers to Socrates that ‹Socrates has phi› attributes phi to Socrates; we also know (same qualification) that, if one couldn't attribute phi to Socrates by saying ‹Socrates has phi›, then "Socrates" would not refer to Socrates. In general, we know that to, the extent that, in virtue of having the form ‹...*E*... ›, a sentence is to the effect that...*O*..., then *E* refers to *O* and that *E* fails to refer to *O* to the extent that this condition isn't satisfied. Given this fact, it follows that we'd be horribly *mis*precisifying the term "reference" were we to say that sentences referred to their truth-values; it also follows that we would not be misprecisifying it *as* horribly were we to say that they referred to the propositions they meant. Give or take some niceties, if *S* is a sentence that has proposition *P* for its meaning, then, in virtue of having the form ‹...*S*... ›, a sentence *S** is to the effect that...*P*...This suggests that, if sentences refer to anything, it is to the propositions they express and, moreover, that it is *not* to their truth-values.

5.0 Davidson on facts

Despite everything thus far said, Donald Davidson (1967) took it for granted that SS is cogent. And he argued that, given the (alleged) fact that SS is cogent, it follows that there is only one fact. His argument is simple. Put the expression "the fact that" in front of each of (1)-(4): an analogue of SS shows that the resulting expressions must co-refer.

Instead of rejecting SS, Davidson concluded, absurdly, that these two facts *are* identical and that, by obvious extensions of this reasoning, there is, at most, only one fact. Having, in his mind, established that there is only one fact, he argued that, for this reason, mental states are not representations. Since there is only one fact, Davidson argued, no two mental states represent different facts. On this basis, Davidson concluded that differences between mental states—between, for example, my perception of the piano in my living room and your perception of the ink-marks in front of you—are not representational differences.

Why is this? Because, says Davidson, representations, supposing them to exist, are representations of facts; and, says Davidson, we've seen that there is only one fact. Therefore, Davidson says, my perception of the piano in my living room doesn't differ in respect of which *fact* it represents from your perception of the ink-marks in front of you. From this Davidson concludes (fallaciously, even relative to his own assumptions), that the one perception is *representationally* just like the other. Thus, the differences between those perceptions have nothing to do with any representational differences that there might be between them. This in its turn means that, if there are any such differences, they are innocuous: in other words, it is not *in terms of* those differences that the palpable differences between those perceptions are to be understood. This, in *its* turn, means either that:

> (i) Neither perception represents anything,

or that:

> (ii) If either perception does represent something, it is irrelevant to how that perception differs from the other perception and, therefore, to what sort of thing that perception is. Which, give or take some unimportant nuances, means that

neither represents anything.

5.1 Evaluating Davidson's contention

Davidson’s position is obviously false. The fact that I am typing this sentence is distinct from the fact that you are reading it. There is more than one fact. What could be more obvious?

Of course, the word "fact" is a vague one; and, surely, it's at least partly for this reason that Davidson's position has been given any credence. In any case, there is no precisification of the term "fact" that validates Davidson's position.

In some contexts, the word "fact" is synonymous with the word "truth." ("The fact that arithmetic is incomplete was unknown to Frege" is synonymous with "the truth that arithmetic is incomplete was unknown to Frege.") Since there are many truths, this interpretation of the word "fact" doesn't validate Davidson's view.

In other contexts, the word "fact" seems to refer to (re)distributions of mass-energy---to the spatiotemporal conditions and changes in virtue of which certain propositions are true and others are false. Let *M* be the mass-energy (re)distribution in virtue of which the proposition *x is now asleep* (where *x* is any sleeping person) is true. Let *N* be the mass-energy distribution in virtue of which *the Earth is now traveling at approximately 17,000 mph with respect to the Sun* is true. $M \neq N$. That's just a fact. Pun intended.

Are there other precisifications of the word fact? Probably. (For example, physical laws may qualify as "facts," at least on one interpretation of that word, even though it's not clear that physical laws are identical either with truths or with (re)distributions of mass-energy). But, apart from Davidson's argument to the contrary, there is no reason to expect any of the remaining precisifications to do a better job of validating Davidson's contention than those we've just considered.

5.2 Evaluating Davidson's argument for his contention

Davidson's argument fails for the same reason *mutatis mutandis* that SS fails. But Davidson's argument involves a blunder of which SS is innocent. For a mental state to be representational is not for there to be some fact that it represents. It is for that mental state to be to the effect that there exists some fact. Suppose that I hallucinate that there is a pink elephant in my living room. (Assume that, in actuality, there is no such elephant.) My hallucination is representational. It isn't like a rock. But there is no elephant *x* such that my hallucination represents some pink elephant *x* such that *x* is in my living room. My hallucination is representational in the sense that it is to the effect that there exists some pink elephant *x* such that *x* is in my living room. For mental state *M* to be representational isn't (necessarily) for there to be some fact that M represents; it is (or can be) for *M* to be to the effect that there exists a fact having such and such characteristics. And the non-existence of such a fact is obviously compatible with the existence of erroneous propositional attitudes to the contrary.

6.0 Neale on a variant of SS

In his (ironically named) book *Facing the Facts,* Stephen Neale puts forth what he believes to be an improvement on SS.[182] Neale's argument involves the following four assumptions:

> G_1: The sentence ‹*Fa*› stands for the same fact as the sentence ‹*a* = the unique thing *x* such that (*x* = *a* and F*x*).› (In this context, "=" is to be taken to mean "is identical with.")
>
> G_2: Any sentence can be put into subject-predicate form.
>
> G_3: What a complex expression refers to is a function of what its parts refer to.
>
> G_4: Definite descriptions are referring terms. "The even prime" is an expression that picks out the number two.

> Assume that for some individuals *a* and *b,* and some properties *G* and *F*, the following are true: ‹*Fa*›, ‹*Gb*›, and ‹*a* ≠ *b*›; and suppose that these sentences stand for facts F_1, F_2, and F_3, respectively.
>
> Bearing these assumptions in mind, consider the following seven sentences.

1. *Fa*. By assumption.
2. *a* ≠ *b*. By assumption.
3. *Gb*. By assumption.
4. *a* = the unique thing *x* such that (*x* = *a* and *Fx*). Follows from 1, by G_1.
5. *a* = the unique thing *x* such that (*x* = a and *x* ≠ *b*). Follows from 2, by G_2.
6. *b* = the unique thing *x* such that (*x* = *b* and *Gx*). Follows from 3, by G_1.
7. *b* = the unique thing *x* such that (*x* = *b* and *x* ≠ *a*). Follows from 2, by G_1.

1 and 4 stand for the same fact, namely F_1. 2, 5, and 7 all stand for the same fact, namely F_2. And 3 and 6 stand for the same fact, namely F_3. ‹The unique thing *x* such that (*x* = *a* and *Fx*)› co-refers with ‹the unique thing *x* such that (*x* = *a* and *x* ≠ *b*)›, since both expressions refer to *a*. Therefore, 5 is what results when the definite description in 4 is replaced with a co-referring definite description. Therefore, 4 and 5 refer to the same fact. Since they refer, respectively, to F_1 and F_2, it follows that F_1 and F_2 are the same fact. ‹The unique thing *x* such that (*x* = *b* and *x* ≠ *a*)› co-refers with ‹the unique thing *x* such that (*x* = *b* and *Gx*)›, since both expression refer to *b*. Therefore, 7 is what results when the definite description in 6 is replaced with a co-referring definite description. Therefore, 6 and 7 refer to the same fact. Since they refer, respectively, to F_3 and F_1, it follows that F_3 and F_1 are the same fact. If G_4 is wrong, then definite descriptions are not referring terms, in which case 5 is not what results when the definite descriptions in 4 is replaced with a co-referring expression, and 7 is what not results when the definite description in 6 is replaced with a co-referring expression. But assuming that G_4 is right, we've established that $F_1 = F_2$. Therefore $F_1 = F_2 = F_3$, and ‹*Fa*› therefore co-refers with ‹*Gb*› and also with ‹*a* ≠ *b*›. So, assuming that definite descriptions are what they appear to be, viz., referring terms, it follows there is only one fact.

This result, if it is in fact true, is astonishing. But G_2 allows us to go further. Let *S* and *S** be *any* two distinct statements that are true. (So *S* and *S** needn't have subject-predicate form.) According to G_2, *S* can be put in subject-predicate form, and so can *S**. Given any existing individual *a*, *S* can be rendered as ‹*a* is an *x* such that *S*›; and given any existing individual *b*, *S** can be rendered as ‹*b* is an *x* such that *S**›. And once that is granted, it follows, by the foregoing argument, that *S* and *S** stand for the same fact. It follows that, given G_1-G_4, any two true sentences stand for

the same fact.

6.1 Evaluating Neale's argument

The problems with this argument are very similar to the problems with Davidson's. Consider line 4 of Neale's argument:

4. ‹*a* = the unique thing *x* such that (*x* = *a* and *Fx*).›

If this sentence is read *attributively* (in other words, if the definite descriptions in them are treated as quantifiers, as opposed to referring terms), then it is logically equivalent with:

1. ‹*Fa*›,

in which case those two sentences can therefore be seen as standing for the same fact. But if 4 is read *referentially* (in other words, if the definite descriptions in them are treated as referring terms, as opposed to quantifiers), then 4 stands for the same fact as:

4_R: ‹*a* = *a*›,

in which case 1 and 4 cannot plausibly be seen as standing for the same fact. For exactly similar reasons, 5, 6, and 7 must be read attributively if there is to be any plausibility to the claim that they stand for the same facts as, respectively, 2, 3, and 4. In general, G_1 is false if 4-7 are read referentially, 4-7 must therefore be read attributively. Bottom line: Neale's argument fails if 4-7 are read referentially. It also fails, for obvious reasons, if they're read attributively.

7.0 A final note concerning SS

SS assumes that:

> (i) Each of the two definite descriptions occurring in it is a singular term.

SS makes the additional assumption that:

> (ii) Reference isn't changed by intersubstituting co-referring terms that *have different senses*.

And, of course, SS assumes that:

> (iii) Referring terms can have both sense and reference (it assumes that the definite descriptions occurring in the second and third sentences have both sense and reference).

(i) is innocuous. Even though, for reasons already discussed, some hold that definite descriptions are *not* singular terms, being instead quantifiers, we could obviously *stipulate* that the definite descriptions in SS are referring terms. We could stipulate that, for any property phi, and for any object *O*, ‹the phi› refers to *O* exactly if *O* uniquely has phi.

But this wouldn't be enough to get SS to go through. For argument's sake, suppose that ‹the phi› refers to *O* exactly if *O* uniquely has phi; and *further* suppose that, in virtue of having the form ‹…the phi…›, a sentence is true exactly if…*O*…In that case, supposing that *O* were a unique phi, ‹Plato thought highly of the phi› would bear a proposition that was true exactly if Plato thought highly of *O*, it being entirely irrelevant what properties (other than being highly thought of by Plato) *O* had, and ‹the phi was a great wrestler› would bear a proposition that was true exactly if *O* was a great wrestler, it being entirely irrelevant what properties (other than being a great wrestler) *O* had.

Thus, supposing that ‹...the phi…› were defined in the way thus proposed, there is some *x* such that *x* = Socrates such that:

> (*) "Plato thought highly of the great philosopher who died

of hemlock poisoning,"

would be true iff:

(**) Plato thought highly of *x*.

Thus, SS not only assumes that ‹the phi› is a singular term. It *also* assumes that, in virtue of having the form ‹…the phi…›, a sentence is true if and only if exactly one thing *O* has phi and…*x*…

But even though each of those assumptions is perfectly defensible, they are incompatible with each other. Unfortunately, the essence of Frege's theory of reference is that *both* assumptions are correct. Thus, that theory is spurious, and so is SS, given that it relies, in a particularly brazen way, on both assumptions. Sadly, the obvious fact that sentences *don't* refer to truth-values wasn't taken to show that Frege's semantics was defective. Instead, it was assumed that Frege's semantics was non-defective that sentences therefore *do* refer to truth-values. For some reason, philosophers haven't been deterred by the fact that there is basis for this contention, apart from Frege's theory. This means that there's no basis for it at all, given that little credence should be given to a semantic theory according to which "snow" refers to an abstract object in "it is true that snow is white." But, in this case as in so many others, philosophers chose to hide their failure to see the obvious beneath a lot of gratuitous logic-chopping.

Chapter 7 Some Puzzles Relating to Language and Logic

1.0 Are there meaningless sentences?

According to some, there are *bona fide* sentences that are meaningless. Among the various confusions underlying this totally erroneous position, three can be cleared up right away:

(i) There are expression-*like* entities that, given their phonetic properties along with facts about the linguistic environments in which they occur, are easily mistaken for *bona fide*expressions. Given what the words flanking it are and what it sounds like, the occurrence of "blerko" in "sometimes a blerko can be a man's best friend" is likely to be mistaken for an actual word. But it has no meaning and is thus no word.

In many lectures of his that I attended (as a student at UCLA), David Kaplan, a well known semanticist, insisted that the opening sentence of Lewis Carroll's *Jabberwocky* (" Twas brillig, and the slithy toves/Did gyre and gimble in the wabe") was a *bona fide sentence.* It is not. It is a sentence-impostor. (Obviously "brillig," "slithy," "tove," etc. are meant to bring to mind the words "brilliant," "slimy," "toad," etc. But those two lines no more token a sentence than a man wearing a Richard Nixon mask is Richard Nixon.)

(ii) According to Carnap (1932), "this stone is thinking of Vienna" is literally meaningless, even though it is a *bona fide* sentence. Carnap is making the familiar mistake of confusing the property of having a meaning worth affirming with the property of having *no* meaning. That stone was *not* thinking of Vienna; so Carnap's sentence was false, and therefore meaningful.

There is obviously a sense in which an utterance of "this stone is thinking about Vienna" is "nonsense." But in *that* sense of the word "nonsense," a sentence's being nonsense consists in its having a meaning, albeit one that has no chance of being correct.

Chomsky (1957) made the very same mistake, though in a

less brazen way, when he famously said that “colorless green ideas sleep furiously” is meaningless but that, since it has the right syntactic properties, is a *bona fide* sentence. On one of its two disambiguations, it says that *there are ideas that are both colorless and green, all of them furious sleepers.* Thus disambiguated, “colorless green ideas sleep furiously” is meaningful, but its meaning isn’t worth affirming. On its other disambiguation, it says that anything that is a member of the class of colorless green things that are ideas is also a member of the class of things that sleep furiously. In other words, it says that nothing that falls into the first class fails to fall into the second class.

Thus disambiguated, that sentence is true but trivial: given that the first class is empty, it follows vacuously that there is no member of the first class that isn’t also a member of the second class. It’s obvious that the first class cannot possibly have any members and therefore cannot possibly have any members in common with any other class. So, thus disambiguated, that sentence affirms a proposition that, being vacuously true and therefore devoid of significant information, is not worth affirming.

Thus, contrary to what Chomsky says, “colorless green ideas sleep furiously” is perfectly meaningful. In fact, it’s hyper-meaningful, since it has two meanings. One of them is a correct (but trivial) proposition; the other is a false (and absurd) proposition. Triviality and falsity are properties of meanings.

So Chomsky’s contention is proven false---and not, contrary to what he seems to think, proven correct---by the very fact that, depending on how it is disambiguated, “colorless green dreams sleeps furiously” is absurd or trivial.

Russell claimed that “quadruplicity drinks procrastination” is a *bona fide* sentence that is meaningless. But that claim is false; the reasons for this are similar to those just given in connection with “colorless green ideas sleep furiously.” First of all, unless “quadruplicity” has a meaning, this sentence of Russell’s is no more a sentence than “drinks procrastination.”

So it must be assumed that “quadruplicity” has a meaning. According to Russell, whose word we can trust here, that word *does* have a meaning; it refers to the property of having four members. (So the smallest set consisting of the numbers 1, 2, 3, and 4 has the property of quadruplicity.) Given this fact, Russell’s supposedly meaningless sentence is quite clearly false, and therefore meaningful. Being abstracts objects, quadruplicity and procrastination neither drink, nor are drunk by, anything. Russell seems to be committing the oft-made mistake of confusing absurdity, which is the property of having a meaning that is logically incapable of being true, with meaninglessness. Russell’s belief that “quadruplicity drinks procrastination” was probably also motivated by his acceptance of the Theory of Types, which we’ll discuss in Section 3.0 of this chapter and which, in agreement with every mathematical and philosophical logician alive, we’ll find to be false.

(iii) Oftentimes, a homonym of an actual word occurs in a context in which no semantic rule assigns any meaning to it and in which it therefore means nothing, even though the context in question is one that is *otherwise* just like a sentence. What occurs in “the absolute is one” is not the word “absolute,” but a mere homonym of that word. For the occurrence of “absolute” in “the absolute is one” to be a word, there would have to be some semantic rule that assigned it a meaning. Although there are semantic rules that assign meaning(s) to homonyms of that occurrence, that occurrence *per se* is assigned no meaning. Anything that is assigned no meaning is *ipso facto* not a linguistic expression. So “the absolute is one” is not a sentence, but a sentence-impostor, at least until it is said what "the absolute" refers to, or is supposed to refer to---at least, that is to say, until the semantic rule for it is identified.

Anything to which no semantic rule assigns a meaning *ipso facto* isn't a linguistic expression. Because the logical positivists overlooked this humdrum fact, they sought some

criterion by which to distinguish "meaningful sentences" (e.g., "the moon is made of cheese") from "meaningless sentences" (e.g., "the absolute is one"). But all of the meaningless so-called sentences that they identified weren't sentences at all; they were sentence-schemata, since they contained undetermined expressions (expressions to which no semantic rule assigned any meaning, at least not as far as occurrences in the sentence in question were concerned) and therefore no more qualified as *bona fide* sentences than does "*x* has property phi."

Thus, in trying to explain the meaningless of pseudo-sentences as “the absolute is one”, the logical positivists were on a fools errant. In that pseudo-sentence, the term “the absolute” is a restricted variable, albeit one whose connotations constrict its range of possible values more than the range of the (almost entirely) non-connotative variable in, say, “x is extended” restricts *its* possible range of values.

2.0 The liar's paradox

We've found elementary blunders in each of the reasons already considered for holding that *bona fide* expressions can be meaningless. But there are other reasons for holding this that don't involve any obvious blunders and demand careful treatment. In the process of considering these reasons, we'll make some headway on some important, open questions in semantics and philosophical logic.

There are sentences that, so initial appearances suggest, are true if they're false and false if they're true.[183] Consider the sentence:

> (LIE_1) "I am now lying."

Given the law of excluded middle ("any proposition must be either true *or* false"), LIE_1 cannot fail to be either true or false. Given the law of non-contradiction ("no proposition can be both true *and* false"), LIE_1 cannot be both true *and* false. So if LIE_1 expresses a proposition, that proposition is both true and false. Therefore, given that no such proposition exists, LIE_1 doesn't express a proposition. But it's very hard to believe that "I am lying" is not a sentence. Exactly similar reasoning holds with respect to "I am lying" and "I am a Cretan and all Cretans are liars," and each one of the infinitely many other sentences that are relevantly similar to those two.

LIE_1 is indeed a sentence. But it's not both true and false; it's just false. LIE_1 is equivalent with:

> (LIE_2) "There already exists some proposition *P* such that *P* is false and such that, by virtue of producing this very utterance, I am affirming *P*."

LIE_2 thus entails that, existing prior to, and therefore independently of, any LIE_2 or any utterance thereof, there is some proposition *P* such that (i) is false and, moreover, such that (ii) is true only if, for some utterance of LIE_2, *P* is true only if that utterance is true. Since *P pre*-exists any such utterance, LIE_2 is not false *and* true; it is simply false. Therefore, LIE_1 is false. So we don't have in LIE_1 a counter-example to our thesis that there are *bona*

*fide*expressions that are meaningless.

2.1 The liar's paradox (continued)

There is another way of dealing with the Liar's paradox, or at least with one version of it. Let UTF be a specific utterance of the sentence:

(TF) "This sentence-token is false."

To what do the first two word-tokens constituting UTF refer? "They refer to UTF", it will be said. "In other words, that occurrence of 'this token' means "the sentence-token I will have produced by virtue of completing this sentence." If only for argument's sake, let's suppose that this is so. In that case, TF is equivalent with:

> (TF*) There exists some proposition P such that there is exactly one sentence-token that I am now uttering such that my utterance affirms P (and is therefore true if, and only if, P is false).[184]

A proposition *is* a truth or a falsehood. So P has been a truth or a falsehood as long as it's existed. (Since—see Chapter 2---propositions don't come into existence or go out of it, this means that it's always existed.) If P pre-dates TF*, then it was true or false before TF* ever came along. In that case, TF* is simply false. For TF* falsely says the truth of P is contingent on the truth of TF*, which is not the case. If P does not pre-date TF*, then TF is simply false; for it starts with the words "there exists a proposition P such that." Even if that proposition pops into existence by the end of the TF*, TF* is false, since it affirmed the existence of some proposition that pre-dated TF*'s completion.

3.0 Russell's Paradox and the Theory of Types

Other sentences that are obviously comparable to "I am now lying" are "this sentence is false" and "I'm a Cretan, and all Cretans are liars." But there exist sentences that, it has been argued, are meaningless and that, unlike those two, are not in any obvious way modeled on "this sentence is false." The best known example of such a sentence is[185]:

(RP) K is a member of itself iff K is not a member of itself,

where K is the class of all classes that are not members of themselves. (If K is a member of itself, then, since K includes only classes that aren't members of themselves, then K is not a member of itself; and if K is not a member of itself, then, since K includes every class that is not a member of itself, K is a member of itself.)[186]

According to Graham Priest, a contemporary logician, RP is meaningful, i.e. there is some proposition that it affirms. According to Priest, since that proposition is both true *and* false, we must reject the law of non-contradiction.[187] In a moment, we'll see why Priest's reasoning is spurious.

Unlike Priest, Bertrand Russell held that the law of non-contradiction is correct and, therefore, that there is no proposition that RP expresses. But Russell has a very different view from us as to *why* RP fails.

Russell's view is embodied in his famous "Theory of Types" ("TT" for short). According to TT, any given sentence belongs to a certain "type"; given any two types, one is higher than the other. The lowest type is Type_1. Sentences of Type_1 concern "individuals." (An individual is anything spatiotemporal and, therefore, anything that isn't a set, property, sentence or proposition. In this context, for the sake of expediency, references to "classes" will be short-hand for references to classes *and* to properties *and* to propositions *and* to sentences.) Sentences of Type_2 concern individuals *and* classes whose members are individuals. (Sentences of Type_2 thus concern *inter alia* sentences of Type_1.) Sentences of Type_3 concern everything with which sentences of Type_2 are concerned *as well as* classes of classes of individuals.[188] In general, for any whole number n, sentences of Type_{n+1}

may concern everything with which sentences of Type $_n$ are concerned and, *unlike* sentences of Type $_n$, sentences of Type $_{n+1}$ may concern classes of Type $_n$.

A consequence of TT is that, if the free variable in “x is a member of class C”, where C is of Type$_m$, is replaced with an expression denoting a class of Type$_m$ or higher, the resulting sentence is meaningless, i.e. there is no proposition that it affirms and it is therefore neither true nor false.

So, according to TT, “C is a member of C” is meaningless, and so is its negation. TT entails that RP is no more meaningful than “blurka derka derk." Given his acceptance of TT, Russell understandably holds it to be a failing of the English language that it permits the construction of sentences like RP.

3.1 Why the theory of types is false

RP is ambiguous; there are two ways of disambiguating it. RP can be taken to be to the same effect as:

> (RP_1) There exists some class *K* such that all and only those classes that are not members of themselves belong to *K*.

Thus disambiguated, RP is false. There is no such class. A contradiction follows from the supposition that there is a class whose membership includes all and only those classes that aren't members of themselves. Therefore, there is no such class.

The fact that no such class exists is to be established by a garden-variety indirect proof, just like the fact that there are no square circles.[189] Let *x* be a square circle. Being a square, *x* has four sides and therefore doesn't have a uniform curvature. Being a circle, *x* has a uniform curvature. Therefore, *x* both does and does not have a uniform curvature. This isn't possible. So *x* is a non-entity. *K* is a non-entity for the same reason *mutatis mutandis.*

There is another way to interpret RP. We could take "*K*" to be defined thus:

> (AI[190]) *Assuming* that there is some class *O* whose members include all and only those classes that are not members of themselves, then, in virtue of having the form ‹…*K*…›, a given sentence *S* is true just in case *O* has…*x*…(i.e., just in case…*O*…); and, if that assumption is false, then "*K*" picks out nothing, and ‹…*K*….› bears no proposition and is thus neither true nor false.

Thus taken, RP doesn't say *anything*. Since there is no class of the requisite kind, "*K*" picks out nothing and, the semantic rule for "*K*" being what it is, RP fails to encode any proposition. Thus disambiguated, KP is like an utterance of:

(TPE) “That pink elephant is coming right at us,”

made in a context where there is no pink elephant. So, thus disambiguated, RP requires neither an abandonment of the laws of classical logic nor an acceptance of TT.

3.2 Two kinds of meaning: type-meaning and token-meaning (revisited)

Here we must revisit an issue discussed in the previous chapter.[191] There seems to be *some* sense in which an utterance of TPE that is made in an elephant-free context is meaningful, notwithstanding that such an utterance bears no proposition. The sense in which such an utterance is meaningful is this: the semantic rules of the language to which it belongs determine what *would* have to be the case for that utterance to be meaningful. The semantic rule corresponding to TPE is this:

> (TPE*) If, in context *C*, there is some pink elephant *E* that is being indicated by some person P_1, and P_1 is indicating no pink elephant other than *E*, then an utterance on P_1's part, directed towards persons $P_2 \ldots P_n$ ($2 \leq n$), of "that pink elephant is heading towards us" is true exactly if *E* is heading towards $P_1 \ldots P_n$. If there is no such elephant, that utterance fails to encode any proposition.

So, when uttered in an elephant-free environment, TPE *is* meaningful. It's meaningful in the sense that it has well-defined truth-*conditions.* Whether those conditions are fulfilled has no bearing on whether TPE is meaningful. Nothing relating to the whereabouts of elephants redounds to the credit or the discredit of the body of rules that constitute the English language.

Similar reasoning holds in connection with RP. RP has truth-conditions. Whether those conditions are fulfilled is irrelevant to whether RP is meaningful. "But there is a difference between RP and TPE," it will be said. "Whereas tokens of TPE *could* be true, tokens of TT cannot possibly be true: the laws of logic prohibit it." Yes---but the laws of logic aren't coded into the semantic rules of English. One can master those rules without being a master-logician; one certainly doesn't need to know whether there exists a class of all classes that aren't members of themselves. So RP, though not proposition-bearing, *is* meaningful.

To conclude: on neither of its disambiguations is RP a counterexample to our contention that all *bona fide*expressions are meaningful. This is another instance of a philosopher failing to distinguish absurdity from

meaninglessness. We discussed other such examples in Chapter 1.

3.3 The theory of types revisited: More problems with it

According to the Theory of Types (TT), no meaningful sentence occupies an order equal to, or lower than, the order occupied by any one of the sentences that it concerns. Since TT itself concerns statements belonging to *every* order, it concerns the order to which it itself belongs and, supposing it correct, is therefore meaningless and, being meaningless, is *not* correct. So TT is not correct if it is correct. Since, therefore, TT is inconsistent with itself, it is false.

Also, TT has consequences too counterintuitive to accept. For example, according to TT, the sentence:

(i) “Jim has three pairs of socks.”

is of a lower order than the sentence:

(ii) “The sentence ‘Jim has three pairs of socks’ is true,"

and (ii) is of a lower type than the sentence:

(iii) “The sentence ‘the sentence “Jim has three pairs of socks” is true’ is true.”

If TT is right, the property that one ascribes to (iii) by describing it as “true” is one that cannot possibly be had by (ii) or (i), and the property one ascribes to (ii) by describing *it* as “true” is one that cannot possibly be had by (i). So, if TT is right, there is no such thing as truth *simpliciter*. There is truth_1, truth_2, etc.

By similar reasoning, if TT is right, the occurrence of “three” in:

(iv) “Jim has three pairs of socks,”

doesn’t have the same meaning as its homonym in (i). (i) is to the effect that:

(a) There are three true instances of the propositional-form *x is a member of the class of pairs of socks had by Jim*.

The occurrence “three” in (i) denotes a property had by classes of individuals. (It denotes the property of being a three-membered class of individuals.) The homonymous occurrence in (iv) denotes a property of classes of classes of individuals. (It denotes the property of being a three-membered class of classes of individuals.) So the property denoted by the one occurrence is distinct from the property denoted by the other. Given how counterintuitive this is, TT must be presumed false.

This presumption is easily verified. TT concerns *all* sentences. It isn’t concerned with sentences belonging to this or that specific order. According to TT, any proposition that concerns all sentences is *ipso facto* meaningless. So if TT is true, then it’s false; therefore, it’s false.

4.0 Vagueness and indeterminacy

The word “indeterminate” has two meanings. The sentence:

(BW) “Bill’s whereabouts are indeterminate,”

can be taken to mean either:

(BW_1) given the available information, it isn’t possible to arrive at a principled judgment as to where Bill is.

or

(BW_2) there is no fact as to where Bill is.

BW_2 cannot possibly be true. At any given point in time, some space-time region *R* is such that Bill is occupying *R*. Of course, half of Bill's body can be in the garage while the other half is in the drive way. But that’s because the region he is occupying is itself half in the garage and half in the driveway; it is not because there is no region that Bill is determinately occupying. The same points *mutatis mutandis* hold of any other proposition to the effect that the world *per se* is indeterminate in any respect. In any case, this is what the present author believes.

But many disagree; and their reasons for doing so seem at first to be good ones. For there certainly *seem* to be cases of objects with vague boundaries. In fact, all spatiotemporal objects, it would seem, have vague boundaries. Given any person *P*, there are many particles *x* such that one cannot non-arbitrarily say that ‹*x* is a part *P*'s body› is true *or* that it's not true. This apparent fact is one that has led many a philosopher to hold that there is objective indeterminacy.[192] But we'll now see that such philosophers haven’t considered all the possible ways of modeling our intuitions.

5.1 Non-epistemic indeterminacy

Smith doesn't have a full head of hair; but there are a few tufts of hair on his head, in between which there is fuzz of varying degrees of thickness and density. Bearing this in mind, consider the statement:

(SB) "Smith is bald."

is SB true or false? Many philosophers, probably most, would hold that:

> (OV)[193] There is no fact of the matter; and that fact has nothing to do with ignorance on our part. Here's what's going on. There exists a certain property *P* that of baldness and a certain person *x*. (*P* = the property of baldness; *x* = Smith.) And *x* doesn't exactly instantiate P, but *x* doesn't exactly fail to instantiate P. There is no fact as to whether the proposition *x has P* is true. It isn't that we don't know whether it's true; it's that there is no fact of the matter.

Here's the problem with OV. For argument's sake, suppose that it is objectively indeterminate whether Smith is bald. Suppose, in other words, that the proposition (not the sentence):[194]

(SB) Smith is bald,

is true.

And, let us also suppose, the reason SB is vague is that it is objectively indeterminate exactly what conditions a given thing must satisfy to be bald.

Given these suppositions, it is *ipso facto* indeterminate what exactly SB entails and what is entailed by it. If it is objectively indeterminate exactly what features a thing must have to be bald, then it is for that very reason indeterminate, for any *x*, exactly what entails, and also what is entailed by, the proposition (not—*nota bene*—the sentence):

(XB) ‹x is bald›,

for any *x*.

Thus, a consequence of the supposition that SB is *objectively* indeterminate is that it is indeterminate what SB entails and also what is entailed by it.

But the very idea of such a proposition is not a coherent one. Propositions are individuated by what they entail. Different entailment relations, different propositions. So until the entailment relations of some sentence-like string of verbiage have been fixed, no proposition has been identified to that verbiage; and all we have is, at most, a sentence-schema.

5.2 An objection to the argument just presented

I have heard it said that:

> (ER)[195] Propositions are *not* individuated by their entailment relations. The proposition that 1 + 1 = 2 obviously isn't identical with the proposition that triangles have three sides. But a given proposition entails/is entailed by the one just in case it entails/is entailed by the other."

Obviously *1 + 1 = 2* and *triangles have three sides* are different propositions. But contrary to what ER assumes, they *don't* coincide in respect of their entailment relations. The *way* [196] in which *for some n, n + 2* follows from *1 + 1 = 2* differs—quite dramatically—from the way (or, rather, any one of the infinitely many possible ways) in which *for some n, n + 2* follows from *triangles have three sides.* A corollary is that, until it is settled what the entailment relations are of a given sentence, it *ipso facto* hasn't been settled which proposition is meant by that sentence. (This doesn't mean that one has to know what all of those relations are. No one knows what *all* of the entailment relations are of "1 + 1 = 2." But, the semantic rules of the English language being what they are, it is indeed fixed what those entailment relations are.)

What a sentence says---or, to be precise, what is affirmed by a non-abortive indicative token of that sentence ---is a function of exactly two things: first, what follows from that sentence-token (supposing it is true); second, *how* those things follow from it (same qualification). Given a sentence-token S, once what it entails and how it entails it is fixed, there is nothing left to be said about what it means. By the same token, if what S entails or how it entails it isn't fixed, it has yet to be fixed what S means. Thus, a sentence such that there is no fact as to what it entails, or as to how it entails what it entails, is a sentence to which no one proposition has been assigned.

Bearing these points in mind, suppose the following. *P* is a true proposition. There are propositions $P_1 \ldots P_n$ such that, for each i, it is objectively indeterminate whether *P* entails P_i. Thus, *P ipso facto* leaves it open whether *P* is true. But if *P* leaves it open whether P_i is true, then *P*

determinately *fails* to entail P_i. That doesn't mean that *P* entails the *negation* of P_i. (*Snow is white* doesn't entail *grass is green* or *grass is not green*. A proposition may be compatible with another proposition *and* with its negation.)

The believer in indeterminacy has but one way of dealing with this argument. He is forced to say that, although there is no fact as to whether *P* entails P_i, nonetheless *P* does not leave it *open* whether P_i is true. But this position is tautologously false.

5.2.1 An objection to the argument just presented (continued)

Here is a similar argument. Suppose there to be some propositions *P** such that (i), for any given proposition, it *is* determinate whether *P** entails that proposition and therefore that it *is* determinate, for any *i*, whether *P** entails P_i but such that (ii), with that qualification, *P** entails a given proposition iff *P* also entails it. Is *P* identical with *P**? It cannot be. For *P** has, whereas *P* lacks, the property of being such that there is a fact as to whether it entails P_i, for some *i*. Thus, *P* is distinct from any proposition *P*% such that *P*%, supposing it true, settles whether or not P_i is true. So *P fails* to settle whether or not P_i is true and thus fails to entail it or its negation. So the statement "*P* entails P_i" is determinately false.

Were there indeterminate statements; i.e., statements such that there was no fact as to whether what they affirmed was true, there would be pairs of statements such that there was no fact as to whether the one entailed the other (or, therefore, whether it entailed its negation). But given any such pair, the one *ipso facto* leaves it open whether the other is true and thus determinately *fails* to entail (the negation of the other), an immediate consequence being that there *is* a fact as to whether the one entails the other. So the concept of a proposition such that there is no fact as to whether it is true is an incoherent one and the same is therefore true of the concept of objective, as opposed to epistemic vagueness.

6.0 The Sorites paradox

Consider the following paradox.

> Paradox #1 (P_1): A person with only one dollar is poor. If a person with n dollars is poor, then so is a person with $n + 1$ dollars. It follows that a person with a billion dollars is poor. But, obviously, a person with a billion dollars is not poor.

For reasons that are of no philosophical importance and that, in any case, will soon become clear, P_1 is an example of what is known as the "Sorites Paradox" (a.k.a., the "Paradox of the Heap," "Sorites" being the plural of the Greek word "soros," which, so I am told, means "heap"). There are three different kinds of Sorites paradoxes, P_1 being an example of only one of those three kinds. We'll now outline solutions to all three kinds, beginning with the kind instantiated by P_1.

Here is an outline of the solution to P_1: Like most predicates, "poor" is a relational expression. Smith may be poor compared to Jones but wealthy compared to Green, or he may be poor *relative* to the standard set by some situation and wealthy *relative* to the standard set by some other situation. If Smith is a millionaire who needs to have a net worth of at least 50 million dollars to become a managing director of company *X*, he is poor relative to the standard of wealth that, given Smith's desire to become a managing director of *X*, is the relevant one. So an utterance of 'Smith is poor' is correct if made in a context in which the relevant standard of wealth is set by Smith's desire to become a managing director of *X*. For, in such a context, a "poor" person is one whose net worth is less than $50,000,000. For that very reason, it follows that, in such a context, *there is* a number *n*-- namely, 49,999,999--- such that an utterance of ‹*x* is poor› is true if *x* has n dollars and false if *x* has $n + 1$ dollars.

Bearing these points in mind, it's clear how to solve each member of the infinitely large class of paradoxes to which P_1 belongs. Consider the following paradox:

> (P_1#) If a person's top running speed is 1 mph, then that person is slow. For any number, if a person whose top running

> speed is *n* mph is slow, then a person whose top running speed is *n* + 1 mph is also slow. Therefore, a person whose top running speed is 75 mph is slow.

A story, structurally similar to the one just told, will make it clear how to solve this paradox. At his best, Smith runs a 100 yard dash in 12 seconds. Smith tries to join track team *X*. In order to qualify for team *X*, one must run a 100 yard dash in no more than 10 seconds. Obviously Smith doesn't make the team. Smith then decides to join team *Y*. In order to qualify for team *Y*, one must run a 100 yard dash in no more than 100 seconds. Smith obviously qualifies for *Y*.

If made in a context in which the relative standard of swiftness is set by Smith's desire to join *X*, an utterance of "Smith is fast" is false. But such an utterance is *true* if made in a context in which the relative standard of swiftness is set by Smith's desire to join *Y*. In a context of either kind, there is some number *n* such that an utterance of ‹*x* is slow› is true if, at his best, *x*'s average speed when running the 100 yard dash is n and such that an utterance of "*x* is fast" is true if *x* is person whose average running speed when running 100 yards is *n* + 1.

In a context of the first kind, *n* is any number such that if a person whose average speed when running a 100 yard dash is n and who, for that reason, runs the 100 yard dash in over 11 seconds but who, if his speed when running the 100 yard dash were *n* + 1, would run it in under 11 seconds. The same thing *mutatis mutandis* holds of contexts of the second kind.

6.1 These points generalized

To generalize these points, we must note a fact about language—more specifically, about predicates.[197] Some predicates are binary. *P* is a binary predicate if the property it expresses is one that cannot be had to varying degrees; it is one that a given thing either has or doesn't have. So "even" (as in "even number") is a binary predicate. Sorites paradoxes always involve non-binary predicates (i.e., predicates that express properties that can be had to varying degrees), e.g., "fat," "tall," "smart." A corollary is that Sorites paradoxes always involve properties that one thing can have to a greater extent than some other thing. One thing can be smarter than another.

Bearing these points in mind, consider the following three sentences:

(i) Smith is smart.
(ii) Smith is *more* smart than Jones.
(iii) Smith is smart*er* than Jones.

(ii) and (iii) are synonyms. In effect, (iii) is just an abbreviation of (ii). Surely, the occurrence of "smart" in (i) has the same meaning (picks out the same property) as the homonymous expression in (ii). This wouldn't be possible if, in (i), the occurrence of "smart" picked out a binary (non-relational) property. This shows that "smart" is *always* relational. This, in its turn, shows that tokens of (i) are elliptical for sentence-tokens that express-relational propositions---sentence-tokens having the form:

(i#) "Smith is smarter than___,"

or perhaps:

(i*) "Smith is smart *enough* for__."

So tokens of "Smith is smart" must be elliptical for sentences such as "Smith is smarter than Jones" and "Smith is smart enough to solve the puzzle." Exactly similar points hold of other non-binary predicates. The meaning of "*x* is fast" is given by some sentence having the form "*x* is fast*er* than *y*" (e.g., "Smith is faster than Jones") or having the form "*x* is fast *enough* to satisfy standard S" (e.g., "that plane is fast enough to make it to

Pittsburgh by 3:00 P.M.”).

6.2 A confusion side-stepped

Here I would like to head off a source of serious confusion. I am not advocating the view that *at the level of semantics* "Smith is smart," "Jones is tall" etc. express relational propositions of the just-mentioned kind. It cannot be assumed that the philosophical insights needed to solve puzzles that turn on facts about language are themselves encoded in the semantic rules of English (or of whatever language happens to be the relevant one). So the just-stated analysis is not a *semantic* analysis; it is not an attempt to identify the semantic rules that assign meaning to sentences of the form "*x* is *P*," where *P* is an arbitrary non-binary predicate.

It makes no sense to say that some number is *more* even than some other number. But this has nothing to do with language; it has to do with the property of being even. The sentence "6 is *more even* than 8," though absurd, does not run afoul of the semantic (or syntactic) rules of English. (Or, if it does, it is only because we have *ex post facto* projected our mathematical insights into the syntactic or semantic rules of English. And were some mathematician to show that evenness came in degrees, it would come to be grammatical---if it isn't already---to say ‹*x* is more even than *y*.›) Clearly, "two is even" is not semantically or otherwise elliptical for some sentences of the form ‹*x* is more even than *y*› or ‹*x* is even enough to satisfy standard *S*.› But unless we implausibly suppose that our knowledge that evenness doesn't come in degrees is coded into the semantic rules of English---unless we implausibly endow the semantic rules of English, which can be mastered by somebody who wrongly believes that evenness *does* come in degrees, with our mathematical knowledge---then ‹*x* is even› is not, at the level of semantics, elliptical for "*x* is more even than *y*" (etc.).

This reasoning *mutatis mutandis* shows that *at the level of semantics* ‹*x* is fast› probably isn't elliptical for ‹*x* is faster than *y*› (etc.). For, were we to say otherwise, we'd be endowing the semantic rules of English with our (extra-linguistic) knowledge of the fact that, indeed, swiftness comes in degrees. Given that those rules can, at least in principle, be mastered by somebody who has the idiotic belief that swiftness does not come in degrees, we mustn't impute to the semantic rules of English our knowledge that swiftness *does* come in degrees; and we must therefore take the position, for which there is a vast amount of independent corroboration, that this knowledge of ours is non-linguistic.

That said, the way to solve P_1 is to take the view that *in some way or other*---probably through pragmatic implicature, not necessarily through literal meaning---tokens of (e.g.) “Bob is smart” are elliptical for tokens of relational sentences of the previously described kind.

6.3 A different kind of Sorites paradox

Let us now move onto a different kind of Sorites paradox. Paradoxes of this kind are much harder to solve than those of the kind just discussed:

> P_2: Let *C* be some cloud and let *m* be some water-molecule dead in center of *C*. Obviously *m* is a part of *C*. But let's suppose that m migrates towards the periphery of *C*. For a while, m will still remain a part of *C*. But if *m* travels, say, two miles away from *C*, then *m* will definitely cease to be a part of *C*.
>
> But where's the cut-off line? Suppose that, for a while, *m* whizzes around *C*'s periphery, sometimes bumping into molecules that are definitely a part of *C*, and never straying from *C*'s periphery---i.e., from what *definitely* either coincides with or falls within *C*'s periphery----by more than an inch. Under those circumstances, is *m* a part of *C*? What about if we continue to suppose that *m* intermittently bumps into *C* but we increase this margin from one inch to five inches? Is *m still* a part of *C*? And how infrequently can *m* collide with molecules that are clearly *C*-constitutive without itself definitely ceasing to be a part of *C*? If such collisions are frequent, we're inclined to consider it a part of *C*, even during those periods when it's not in contact with molecules that are clearly *C*-constitutive (provided that, during such periods, it doesn't veer off by *too* great a distance). By the same token, if such collisions are infrequent, we're inclined *not* to consider *m* a constituent of *C*, even if, in between such collisions, it remains fairly close to *C*. (The more frequent the collisions, the more *m* can travel away from *C without* ceasing to be a constituent of *C*.)
>
> But, once again, where is the cut-off line? Is it half an inch? Two inches? A foot? What is it? Any answer seems arbitrary. If I say: "it's exactly 3 inches", I'm making a stipulation; I'm not identifying a pre-existing fact. (Technically, the right question is: where are the cut-off lines? For, whether or not m is a member of C seems to depend *both* on its average proximity *and* on the frequency with which it comes into contact with *bona fide*C-constituents. But, for simplicity's sake, I've chosen to abstract

from this fact.) So here, quite indisputably, we have a case of indeterminacy. But, more interestingly, this indeterminacy has nothing to do with the representations of the world that we, or creatures of any other kind, have of the world. There could (and probably would) be clouds even if there were no sentient beings; and, in the absence of such beings, there would, given just about any cloud *C*, be molecules that were definitely parts of *C* and others that definitely weren't parts of *C and*, what is relevant, molecules *x* such that *x is a constituent* of *C* wouldn't exactly be true and wouldn't exactly be false---or, to make an equivalent point, such that *x's constituency in C* would not itself exactly be a constituent of the universe and also wouldn't exactly fail to be a constituent of the universe.

6.3.1 A correct conception of thinghood the key to solving P_2

The solution to P_2 is similar to the solution to P_1. But there are some significant differences; and to solve P_2, we must first make some substantive claims as to what is meant by the words such as “thing” and “object.” (In this context, those terms are to refer only entities such as vases, rocks, windows, as opposed to *abstracta* such as numbers or possibilities.) For the reasons given in Sections 5.2-5.2.1 and now to be developed, I don’t consider it an option to say that the world *per se* is indeterminate. Whatever the word "cloud" means, we gave it that meaning. So the corresponding cannot possibly contain the information needed to answer substantive questions concerning cloud-composition, given that we didn't built such knowledge into it.

Semantic rules cannot know more than we do. So it's understandable, if not inevitable, that there should be arrangements of molecules that determinately don't deserve to be characterized either as "clouds" or as "non-clouds." and it's understandable, if not inevitable, the word "cloud" should fail to pick out *any* one property.

We cannot expect our linguistic conventions to "cut nature at the joints." to use Francis Bacon's expression; we cannot expect them to be hewed to exactly those properties in terms of which the constitution of the world is to be delineated. A slight detour through legal theory will help make this clear.

Laws created 200 years ago necessarily fail to address issues relating to the disposal of motor-oil. There's nothing mysterious about this, and it's obviously no reason for saying that the world is intrinsically vague. The same holds of semantic rules that were created over 200 years ago. And semantic rules, like laws, must be updated if they are to do what we want them do. So it isn't hard to make sense of the idea that semantic rules should sometimes be silent.

But it *is* hard to make sense of the idea that properties should be indeterminate. What does it mean to say that the *property* of being a cloud is indeterminate? To which property are we referring? "The one had by those billowy things in the sky." Yes, but there are other objects that this person is not currently indicating. So, when confronted with some new object, by what criterion do we decide whether it is of the same ilk as those billowy things? Even if we could see the microstructural properties of those objects, we wouldn't necessarily know in virtue of which facts about their microstructures

they fell into the same category as those billowy things.

Until we can answer these questions, we don't have an answer to our original question, namely: *which* property is the relevant one? The information we have limits what a viable answer to that question could be. But it still leaves us with an infinitely large class of distinct properties to choose from.

To sum up: attempts to identify properties that are objectively indeterminate can be interpreted as cases where *no* one property has been identified, suggesting that so-called cases of objective indeterminacy are nothing more than projections into the external world of the platitude that open scientific questions are not to be solved by studying semantic rules.

6.3.2 A correct conception of thinghood the key to solving P_2 (continued)

In any case, P_2 arises because we think about things in the wrong way. We think of things as being more thing-like than they really are! Right now I'm looking at the pen on my desk. Let *x* be one half of my pen (the segment that includes the pen's tip and whose length is half the pen's), and let *y* be some region of the table immediately adjacent to *x*. Finally, let *Z* be the "entity", for lack of a better word, whose only two parts are *x* and *y*. *Z* is not a 'thing.' If we were to list all the things in the world, my pen would be on the list, and so would the table on which it rests. But *Z* wouldn't be on the list. Why not? Because *Z* lacks the dynamic cohesiveness, and concomitant predictive and explanatory usefulness, that is had by the pen or the table or any other *bona fide*object.

Of course, tables don't last forever. They can be broken and burnt; and they all come to an end, in one way or another. But it is never *by virtue of* being a table that a thing ceases to exist. *Qua* table, a thing can be depended on and its behavior predicted. The fact that poorly constructed tables are not thus reliable only confirms this point; for, in virtue of being poorly constructed, and for that reason lacking causal and explanatory integrity, a thing *fails* as a table. This only confirms that it is never *as a table* that a thing lacks causal or predictive usefulness.

These points *mutatis mutandis* hold of *all* objects. Things are *ipso facto* prediction- and explanation facilitators. One thing can be more prediction- and explanation-conducive than another. Because of its relatively high degree of causal integrity, a house made of bricks and mortar is more explanation- and prediction-conducive than a house of cards. We are independently inclined to say that, other things being equal, brick house is more of a thing, and thus more "thingy," than a house of cards. We now know what it is that we're saying when we say this. We're saying that a brick house *ceteris paribus* has more structural integrity a house of cards. Thus, our analysis that things are prediction-/explanation-facilitators is consistent with our pre-theoretic intuition that (to switch examples) clouds, though things in some sense of the word, are less thingy than houses.

6.3.3 A consequence of our analysis

One consequence of our analysis is that a given aggregate of molecules may be a thing in one respect and fail to be a thing in some other respect. In a context where the interactions holding among molecules m_1…..m_n facilitate prediction and explanation, m_1….m_n constitute a thing. In a different explanatory context, one where they don't have the aforementioned virtues, they don't constitute a thing. A galaxy is a vast object. But a galaxy does, in some contexts, function as a single object; it has the requisite causal integrity and predictive fertility that, when trying to answer certain questions, one does well to consider the various microparticles of which a galaxy consists as constituting a single object. In other contexts, one is ill-advised to do this, for the same reason *mutatis mutandis.*

To be sure, we believe that some aggregates of molecules (e.g., vases, rocks, pens) are *just* things; we believe, in other words, that their having the status of thing isn't relative to the dialectical context. This viewpoint reflects prejudices that, in their turn, reflect idiosyncratic facts about our lives. Supposing that m_1…..m_n are the molecules composing the vase on your coffee table, if we were much smaller, so that vases were to us what, in actuality, galaxies are to us now, whether m_1…..m_n constituted a single object would be a function of the explanatory context. Similarly, if we were much bigger and the rate at which we uploaded and processed information were much slower, we might see galaxies in much the same way that, in actuality, we see vases—we might see them (wrongly) as being as unconditionally thingy, since, relative to what we'd need to know under such circumstances, it would (given certain additional assumptions as to what our interests and needs would be, were we such slow-thinking giants) only be in hypothetical or, at least, rather exotic contexts that they lacked the properties requisite for thing-status.

6.3.4 The solution to P_2

These facts about the concept of thinghood help us solve P_2. It's meaningless to ask "how far can m stray from *C*'s periphery---i.e., from the outermost layer of molecules that are *definitely* constitutive of *C*---without ceasing to belong to *C*?" The reason it's meaningless is that whether *m* is, or is not, a part of *C* depends on the explanatory context. Relative to certain questions about *C*, *m* would definitely fail to qualify as part of *C* were it so much as one tenth of an inch outside of the aggregate of molecules that *definitely* qualified as being identical with, or very close to, *C*'s surface. (We'll henceforth refer to that molecule-aggregate as "*CS*"—short for "*C*'s surface.") In other contexts, *m* would definitely be a part of *C* if it were separated by a distance of 10 inches (or less) from *CS*. (Remember that, in this context, we're assuming, as to simplify the discussion, that *m*'s distance from *S* is the only thing that matters, so far as *m*'s being a constituent of *C* is concerned. In all likelihood, other things would matter, e.g. the manner in which *m* was moving, the frequency with which it collided with indisputably *C*-constitutive molecules and the manner in which such collisions occurred, etc. But analogues of the points we're making hold with respect to the questions "how *often* must *m* collide with molecules that are clearly *C*-constitutive?", "with much force, and in what exact manner, must *m* collide with such molecules", and so on.)

In contexts of either of these two kinds, there would be some number *n* such that, if *m*'s distance from *S* were *n* or less, *m* would determinately belong to *C* and such that if *m*'s distance from *S* were $n + 1$, *m* would definitely fail to qualify as a member of *C*. In contexts of the third kind, this would not be the case. But that is because, in such contexts, "*C*" would have no definite referent. It is *not* because, owing to some "objective" indeterminacy---some gap or blank in the world---there was no fact as to whether or not *m* was a part of *C*. The real reason would be as follows.

6.3.5 An objection to our solution

“But aren’t there contexts," will asked, “where, relative to the question whether *m* was or was not a constituent of *C*, it wouldn’t matter by exactly how much m was separated from *CS*? And, supposing that such contexts exist, which they presumably do, doesn’t that show that there is objective indeterminacy?”

Each of these questions is to be answered with a “no.” Let m be some molecule that is on the periphery of some cloud C. We'll draw a blank if, without having any other questions in mind, we ask ‹is m a part of C?›. The reason: the question is meaningless. Questions of the form: ‹is x a part of y?› are seldom to be decided on strictly geographical grounds. Such far as such questions *seem* to be of a strictly geographical nature, a thing's causal liaisons to other things---what it affects, and is affected by, and in what manner---is obviously dependent on where it is. But, when we *think* that of the form ‹is x a part of y?› is a strictly geographical one, the reason is almost certainly that, given where x is, it's clear that each of many different questions as to x's dynamic relations to other entities is inevitably to be answered in the affirmative. It is only in very artificial contexts that the question ‹is x a part of y?› is to be answered on strictly geographical grounds. And given such a context, the question is meaningless unless the relevant geographical criterion is specified. But once it's specified, the question clearly admits of an affirmative or negative response.

Whether x is a part of y depends what sort of *organization* y is and on how x contributes to that organization. My body is an organization. That's why an object lodged deeply in my body isn't necessarily a part of it. A penny that I've accidentally swallowed isn't a part of my body; it doesn't contribute to that organization in the right way. The United States is an organization. That's why a non-American can be in Delaware and an American can be in Nepal. He does little for the people of Nepal, and *vice versa*, and he expects little from them, and *vice versa*. Any responsibilities that he has towards them, and any expectations that he has of them, are likely to be of an *ad hoc* nature. Their laws don't protect him; his soldiers don't protect them. It's somebody in Washington who signs his paycheck and decides whether, this next year, he'll still have a job and, if so, what job will be.

We naturally think of clouds as paradigm cases of things that are *not* organizations. So, given some molecule m and some cloud C, we assume that

the question ‹is m a part of C?› isn't remotely comparable to questions, such as "is Smith a U.S. citizen?" and "is Smith's right leg a very realistic looking prosthetic (in which case, it is not a part of his body) or is it actually his leg (in which case, it *is* a *bona fide*body part of his)?," the answers to which clearly depend, not just on the relative locations of two objects, but on the ways in which the objects in question interact. We tend to believe that:

> (PP)[198] Instances of such relationships involve more than one object's being a part of another. Pure parthood is a strictly spatial relation. If x's being a so-called "part" of y involves x's being part of an organized system of entities, then x's relation to y isn't one of mere parthood. Consequently, an organized systems of entities is *ipso facto* such that its constituents cannot possibly bear that relation to it. Thus, anything whose constituents *do* bear this relationship to them is *ipso facto* nothing more than a disorganized heap or aggregate of entities.

But PP is antithetical to the truth. Not only *can* organizations have parts, contrary to what PP alleges: they are the *only* things that have parts. The relation of parthood is *always* to be understood in organizational terms. Whether object x can fit into organization y often depends on where x is in relation to y. (Hence the illusion that parthood is sometimes a purely spatial relation.) But it's only if x's bearing a certain spatial relation to y suffices for its being appropriately integrated into y that x's location relative to y suffices for its being a part of y.

So supposing that R is the space-time region occupied by some cloud, and some golf ball entered that cloud and, by some miracle simply hovered there for the duration of the cloud's existence, it would never cease to be a foreign object. The reason: the golf ball's structure doesn't permit it to behave in ways in virtue of which R's occupant interacts with other objects in distinctively cloud-like ways.

Thus, even clouds aren't mere heaps of molecules. *Nothing* is a mere heap. Any group of objects that is organized in one respect is disorganized in some other. The very people who compose a highly organized law firm may compose a highly disorganized comedy troupe. The same particles that make up a highly organized work of art may make up a highly disorganized floorboard.

There's nothing paradoxical about this. From a strictly syntactic standpoint, "organized" is a one-place predicate and thus denotes a property, not a relation. But, from a logical standpoint, "organized" a two place predicate and thus denotes a relation, not a property. For Smith to be organized when it comes to his responsibilities to his employer and disorganized when it comes to his responsibilities to his children is for Smith to bear some relation to the desiderata one must satisfy to be a good employee that he *fails* to bear to the desiderata that one must satisfy to be a good father.

Before closing the argument, let's sum up. A moment ago, we saw that x's being a part of y consists in y's being a system of into which x has been integrated. But we just saw that "organized" is a relative term. A group of objects cannot *just* be organized; it must organized in this or that respect, a corollary being that a group of objects that is organized in one respect is necessarily disorganized in some other.

So before we ask whether x is a part of y, there are a few questions that we must answer:

> (i) What kind of organization are we taking y to be?
>
> (ii) What conditions a given object must satisfy to belong to such an organization? How must a given object interact with a group of objects composing such an organization if it is to be one of them?
>
> (iii) Does x interact with the objects composing y in the requisite ways?

These questions may be hard to answer. But they're answerable, at least in principle, and once they're answered there cannot possibly remain any reason to hold that there is no fact as to whether x is a part of y.

6.4 A third kind of Sorites Paradox

Let us now discuss, and attempt to solve, the so-called "Paradox of the Heap."[199] Here is that paradox:

> P_3: A single grain of sand is not a heap of sand. If *n* grains of sand don't jointly constitute a heap, then neither do *n* + 1 grains of sand. In other words, adding a grain of sand to a non-heap won't yield a heap. Therefore, a billion grains of sand don't constitute a heap.

As it stands, P_3 isn't really a paradox. First of all, a billion grains of sand, even when taken jointly, *don't* necessarily constitute a heap. If a million grains of sand are dispersed through space---if they are 'scattered'---they don't constitute a heap. So for a given multiplicity of grains of sand to constitute a heap, any given grain must be (relatively[200]) contiguous with at least one other member of that multiplicity. But even though, for a multiplicity of grains of sand to form a heap, it is necessary that they fulfill that condition, it isn't sufficient. For a billion grains of sand that are arranged in a straight line, each contiguous with its neighbors (or, if it's the first or last grain, with its one neighbor) isn't a heap. For a multiplicity of grains of sand to qualify as a heap, it is necessary and (ultimately) sufficient that they satisfy the following requirements:

> (i) each member of that heap must be contiguous with at least one other such member;
>
> (ii) taken jointly, those grains of sand must form a 3-D entity;

and

> (iii) the entity in question cannot be a *structure*, at least not

in the same sense as a building or a ship in a bottle.

(i) and (ii) are self-explanatory, but (iii) requires clarification. A sand-*castle* is not a heap of sand. What is the relevant difference? “Sand-castles are products of human artifice, whereas heaps of sand are not.”

First of all, heaps of sand are often products of human artifice; second of all, there are structures (e.g., plants) that are not the products of artifice that are definitely heaps and wouldn’t be heaps if, other things being equal, they were made of sand. The difference between a heap of sand and a sand-castle seems to lie in the fact that there are quite strict conditions that a grain of sand must satisfy to become part of a sand-castle, whereas there are no conditions, apart from some very generic ones, that a grain of sand must satisfy to become part of a heap of sand. To become part of a given sand-castle, a given grain of sand must be affixed to the sand-castle in such a way that it either (a) provides extra support for the walkway connecting the two towers or (b) completes some otherwise incomplete decoration on the castle, etc. But to become part of a heap of sand, a given grain of sand need only be adjacent with one or more of that heap’s existing constituents.

These points make it clear how to solve a paradox that is similar to P_3 and that is to be solved in much the same way:

> (P_3#) A single grain of sand is not a heap. If *n* grains of sand, no matter how they are arranged, don’t jointly constitute a heap, then adding another grain to those grains won’t yield a heap, no matter how that extra grain is added to that heap.

What makes it seem that (P_3#) is hard to solve is that the word “heap” hasn’t been defined. I don’t mean that it hasn’t been ***precisified***[[201]], but that it hasn’t been disambiguated. If a “heap” refers to a *structure*, as opposed to an amorphous mound, of a certain kind, then, depending on what sort of structure that is, adding one more grain of sand to a structure consisting of n grains of sand quite definitely will turn a non-heap into a heap.

If by a “heap” we mean an amorphous mound of some kind, then, given a few fairly innocuous points about language that we’ve already made, we’ve

already solved P_3. Let us (speaking as logicians, not as semanticists) regard "heap" as a noun form of the adjective "heapy." There are obviously degrees of heapiness---just as there are degrees of intelligence, tallness, etc. Whether a given mound of sand is "heapy" is relative to the operative standard. Let A and B consist of n and $n + 1$ grains of sand, respectively, but otherwise be qualitatively identical. To an ant whose objective is make it over mound A, A may not be heapy---meaning that the ant can make it over. But B may be heapy to that same ant, given that, because B contains that extra grain of sand (as to whose position in B I am making definite assumptions, clearly), B is not surmountable. So relative to the ant's objectives (and also, perhaps, relative to the ways it can think of surmounting a heap), B is, whereas A is not, heapy.

6.4.1 Summing up

P_3 is dealt with in two steps.

(1) Replace “heap” with the corresponding adjective.

(2) See that adjective as having (hidden) relational structure.

7.0 Zeno's Paradox

In order to travel a distance of n yards, it is necessary to travel a distance of n/10 yards. In order to travel a distance of n/10 yards, it is necessary to travel a distance of n/100 yards. In general, one cannot travel any distance at all without already having traveled some smaller distance.

Consequently,

(i) Any given motion presupposes a prior motion.

A consequence of (i) is:

(ii) There is no such thing as a stationary object's *beginning* to move,

Another consequence of (i) is:

(iii) There is no such thing as an object's ceasing to be in a given place.

A consequence of (iii) is

(iv) Nothing moves.

Therefore nothing moves.
But they obviously do!
How to solve this paradox?

7.1. Eugene Mills' solution to Zeno's paradox

For any possible universe U, there is some distance D such that nothing in U can move at all without covering a distance of at least D.

Explanation: Zeno has established that *if* any given motion presupposes another motion, then nothing can even begin to leave the place that it is occupying at a given time. Therefore, (i) must be rejected.

Assuming that (i) is false, it is not possible for there to no minimum, short of nullity, to the amount by which a given object moves. Therefore, in any world where motion is possible, there is some finite distance such that anything that moves in that world moves by at least that distance.

7 .1.1 Problems with Mills' solution

If Mills is right, then any given universe U consists of discrete cells such that, for any occupant x of U, x moves by vanishing from one cell and then appearing in another. x cannot *move* from one cell to the next, since there would in that case *not* be a minimum distance that x had to travel in order to travel at all.

A consequence of Mills' position is that *nothing* moves. The assertion that

> (i) x spontaneously appears in one cell and spontaneously reappears in an adjacent cell

Is equivalent, logically and evidentially, with the assertion that

> (ii) x spontaneously ceases to exist and an x-*like* object that is not identical with x spontaneously appears in a cell adjacent to that occupied by x.

If you are obliterated and some person who is exactly like you appears a moment later right next to where you were standing, that person isn't you. Let you_2 be that person. Given that you_2 came into existence spontaneously, there are no lines of transmission linking him to the past and thus no lines of transmission linking you_2 to you. Therefore, you_2 is not you; and you_n $(n>2)$ isn't identical with either you or with you_i $(1\leq i\leq n)$ any one of you-you $_{n-1}$. If n is a sufficiently high number and there were a sufficiently small delay between you_i's disappearance and you_{i+1}'s appearance, it would certainly *appear* that you had moved from one place to another. But that is obviously irrelevant.

In a Mill-universe, nothing has a past and, consequently, nothing has a future: everything that exists, exists for an instant and then vanishes. There would therefore be no possibility of mutual causal influence among occupants of a Mill-universe.

But such a universe is no universe at all. Each of its occupants is isolated from each other such occupant. Therefore, each such occupant is in the very condition that it would be in if it were in a universe all by itself. Therefore, a

Mill-universe is a non-universe.

Also, relative spatial and temporal position are to be understood in causal terms. Events are simultaneous if neither can affect the other; otherwise non-simultaneous. Since there is no causation in a given Mill-universe U, no U-occupant x can have either a spatial or a temporal position relative to any other U-occupant y. Thus, a Mill-universe is one in which there are neither places nor times.

In conclusion, Mill has given us a heap of absurdities, not a solution.

7.2 Bertrand Russell's solution to Zeno's paradox

In general, the sum of infinitely many positive numbers may be finite (e.g. 1+½+¼+…=2), and any finite number can be represented as such a sum. Therefore, the sum of infinitely many non-null motions may be a finite motion, and any given finite motion can be represented as such a sum. It is therefore no more of a mystery that any given motion should follow an infinite series of motions than it is that 2 should equal 1+½+¼+…

7.2.1 Problems with Russell's solution

This is not a good solution. Zeno is asking how a given object can take that *first* step. He is asking how an object that moves 2 inches can move inch, and then another ½ inch, and then another ¼ inch, and so on. Russell hasn't answered that question. He has only made the irrelevant point that *if* a given object can move that first inch, *then* it can, after an infinite series of further motions, move a second inch.

"The first step is the hardest," said Mao Tse Tung. Once an object moves that first inch, all of the remaining inches are a foregone conclusion. This is the conceit underlying both Russell's solution and Mills'. But each of these solutions merely begs the question, given that the very thing Zeno is asking is how that first step can be taken.

7.3 The solution to Zeno's paradox

All position is relative position. Therefore, changes in position are changes in relative position. If x is the only occupant of a universe U, it cannot be said that x moves or *a fortiori* that x moves by a given amount. For there is nothing in relation to which x's position can change.

Nor can it be said that x exists at a given time or for a certain period of time. For there is nothing that x either follows or precedes or with which it is simultaneous.

Therefore, the universe occupied by x comprises *no* spatiotemporal positions, except such as are internal to x's proper parts. So far as there are relative positions or changes therein, they are strictly x-internal.

x can equally be thought of as the sole occupant of U or it can be thought of as U itself, it being x's proper parts to which relative positions and therefore changes in relative position can be ascribed and it therefore being x's proper parts to which positions of any kind can be ascribed.

The sole occupant of a universe *is* that universe. Contrariwise, a space-time manifold has no existence independently of its occupants.

Position and change therein are *multiply* relative. Given only two objects, x_1 and x_2, it can be said that x_1 and x_2 are now further from each other, or closer to each other, than they were previously. But it cannot be said that either of them has changed its relative position by this or that specific amount. This is because, just as position is relative position, so distance is *relative* distance. If, in addition to x_1 and x_2, there is a third object x_3, the distance between x_1 and x_2 can be compared with the distance between x_1 and x_3, and it can be said that the one distance either equals or exceeds the other. But unless there is a fourth object x_4, it cannot be said by *how* much the one distance exceeds the other.

The more objects there are in a given universe, the greater the number of distances with which a given distance can be compared. The more such distances, the higher degree the of precision with which a given distance can be meaningfully described.

Therefore, if $x_1 \ldots x_n$ are the occupants of a given universe, changes in x_i's position are capable of being small to the extent that n is large. It is the value of n that sets the metric. What we think of as a case of motion on the part of a

given object x_i is really a change undergone by $x_1 \ldots x_n$ collectively. Seemingly system-internal changes are systemic changes. Therefore, to say that motion is continuous is to make the correct, if trivial, point that the class of changes capable of being undergone by a universe whose sole occupants are $x_1 \ldots x_n$ is, for any number m, a proper subset of the class of changes capable of being undergone by an otherwise identical universe whose occupants are $x_1 \ldots x_n \ldots x_{n+m}$.

So when it is said that x_i must move ½ inch in order to move an inch, and ¼ inch in order to move a ½ inch, and so on, the point being made is this:

> (1) Given any change, however minute, in an object's position, if the universe were better populated and otherwise more heterogeneous than it currently is, it would be possible for such a change to be even more minute.

Given that *x_i moved .n inches* is really an abbreviated way of describing a change collectively undergone by *all* of the universe's occupants. Therefore, the meaning of (1) is given by the truth that:

> (2) No matter how well populated and otherwise heterogeneous the universe is, the universe is capable of being even *more* populous and otherwise heterogeneous; and for this reason the class of possible states of the universe is a proper subset of the class of possible states of a more heterogeneous but otherwise identical universe.

Let us consider the following two statements in light of (2)

> (a) Any given motion consists of other, smaller motion; there is no minimal unit of distance

Given (2), it follows that (a) has nothing to do with the structure of motion, its real meaning being given by

(b) If K is the class of possible states of the universe as it currently is and K* is the class of possible states of an otherwise identical universe that is more heterogeneous than others, K is a proper subset of K*.

There is nothing paradoxical about (b). It is a truism, in fact. And there is no analogue of Zeno's paradox that holds in connection with (b).

Systemic changes are too complex for us to understand with any precision. It is only by conceiving of them as system-internal changes that we can have any understanding of them. We should therefore expect contradictions to arise if take such approximations as *bona fide*truths.

Chapter 8 Do We Think in Words?

1.0 Two different ways that knowledge of a language might enhance intelligence

It is obvious that knowledge of a language makes one more intelligent in at least some respects. Consider the thought that authoritarian governments tend to be less responsive to the needs of their constituents than democratic governments. Could one grasp that thought without knowing some language or other? We have, I think, at least *some* inclination to say "no." And, more generally, we are reluctant to regard creatures that cannot interact with reality through a language, or in any case through some sort of symbolic medium or other, as being capable of grasping, let alone delineating the consequences of, abstract principles or therefore of thinking about reality in a *discursive* manner.

My own view, as we'll see shortly, is that this is false—that whatever one can grasp *through* language, one must be able to grasp *independently* of language. But this, I grant, is a counterintuitive view; and, speaking personally, it was only after trying to defend the negation of that view, and finding that it was impossible to do so, that I came to accept it. Still, our untutored intuitions do, I think, incline us to hold that, but for knowledge of a language (of some kind or other), a creature would be incapable of thinking discursively about the world.

But untutored intuitions are not always probative. And sometimes they are inconsistent with one another, this being one of those cases. For, while we are reluctant to impute discursive thought to the language-less, we *also* strongly feel that we sometimes have thoughts of considerable sophistication that we don't know how to put into words; and, if that is right, then it seems to follow that not all discursive thought is language-dependent.

The purpose of this chapter is to provide a brisk but relatively rigorous answer to the question: Do we think in language?

2.0 Two different versions of the thesis that we think in language

Do we think in language? No. But, before we can say why, there are two very different versions of the view that we think in language.

Version #1: English speakers think in English; Albanian speakers think in Albanian; etc. In general, people think in the *natural* languages that they learn.[202] (A "natural language" is one that organically arises through human interactions. So English, Spanish, Arabic, Chinese, etc. are natural languages. A non-natural or "artificial" language would be one that people invent—for example, computer-programmers invent languages to serve their special needs, and so do mathematicians and philosophers. Philosophers invent special notations because natural languages often obscure important features of the logical structure of the meanings that their sentences express.)

Version #2: We think in an innately known code, a "language of thought." So although we think in sentences, those sentences don't belong to English or Spanish (etc.) but rather to some system of symbolism that is hard-wired into us from birth.[203]

3.0 Why version #1 is wrong

Do English speakers think in English, Spanish speakers in Spanish, etc.? No. (We will find that some of the arguments against Version #1 also apply to Version #2.)

3.1 First reason

We can obviously have thoughts that we are incapable of putting into words. This shows that we grasp those thoughts in some way *other* than by grasping the English sentences that express them. Also, there are occasions where, although we *can* put into words what we are thinking, doing so requires great effort: we have to think long and hard about which words to use. This shows that thinking is often a *pre*requisite to using language, which in turn shows that at least some thinking is non-linguistic.

Astonishingly, many authors (e.g., Ludwig Wittgenstein, Simon Blackburn, John McDowell) *deny* that we know anything that we can't readily articulate. "If you can't say it, you don't know it," they say. This presupposes that what we believe is expressed *only* in the words we utter. But that's not the case. A given person's beliefs can be expressed in many ways, and many of them don't depend on that person's being able to use words. A person's knowledge of the differences between *x* and *y* may be expressed in that person's drawings of *x* and *y*; or it may be expressed in that person's movements while he is in the vicinity of *x* and *y*; or it may be expressed in some other non-linguistic manner.

The way that authors such as Blackburn, McDowell, and Wittgenstein respond to this is to say that no *bona fide* knowledge is involved in my painting things accurately—that what is involved is a kind of pseudo-knowledge. This is obviously false. It also overlooks the fact that, but for this "pseudo-knowledge," we couldn't know anything about language, since it is only through our ability to differentiate noises, ink-marks, etc., on the basis of their perceptible properties that we can learn, or use, language.

3.2 Second reason

Some kinds of information (e.g., visual information) seem inherently incapable of being fully expressed by sentences of English or of any other natural language. Also, sentences have very different structures from pictures, even when the sentences and the pictures overlap in content. A visual perception of a man wearing a leather jacket has a very different structure from any sentence that describes that experience. For example, such a sentence decomposes into a finite number of distinct parts (words), whereas the sense-perception does not so decompose. (Sentences are ***digital*** structures, meaning that they consist of discrete entities. Sense-perceptions are ***analogue*** structures, meaning that they don't.)

Also, any sentence must be interpreted to be understood. To understand the sentence "there is a man wearing a leather jacket over there," you must know the relevant semantic rule (i.e., you must know the rules that assign meaning to that sentence). But you don't have to interpret a visual perception, at least not in the same way. Perceptions are "self-interpreting" in a way that sentences are not.[204]

3.3 Third reason

In order to learn a language, one must already be able to think. To learn English is to learn that certain sounds have certain meanings; and in order to be able to do this, you already have to be able to synthesize information—to make judgments, on the basis of experience, as to what sounds have what meanings. So knowledge of a language *pre*supposes the ability to think.

3.4 Fourth reason

Different sentences can have the same meaning, and a single sentence can have multiple meanings. What you think when you think that Bob loves Mary is identical with what you think when you think that Mary is loved by Bob. But the sentence “Bob loves Mary” is different from the sentence “Mary is loved by Bob,” even though they mean the same thing. So if you thought in English sentences, then the belief you’d have by virtue of accepting the sentence “Bob loves Mary” would be totally different from the belief you’d have by virtue of accepting the sentence “Mary is loved by Bob.” But the belief you have in virtue of accepting the one sentence is identical with the belief you have in virtue of accepting the other.

Also, if you thought in language, you couldn’t disambiguate ambiguous sentences. If your thinking *Bob is dumb (i.e., unintelligent)* were identical with there being an occurrence in your mind of the sentence “Bob is dumb,” then your thinking *Bob is dumb (i.e., he can’t speak)* would be identical with there being an occurrence in your mind of that same sentence going off in your head, and those thoughts would therefore be identical. But they are not identical. So your thinking *Bob is dumb (i.e., unintelligent)* is not identical with your thinking “Bob is dumb.”[205]

3.5 Fifth reason (similar to reason #4)

People who speak different languages can have the same thoughts. Monolingual Spanish speakers believe that 1 + 1 = 2, and so do monolingual English speakers. But English and Spanish speakers use different sentences to express that truth, showing that thinking that truth cannot be identical with thinking a sentence of English or Spanish or, by obvious extensions of this line of thought, any other language.

3.6 Sixth reason

Understanding sentences involves thought. You must be able to think to understand what is meant by “if snow is cold, then snow is not hot.” This is because, in order to understand that sentence, you must be able to pair it off with the right meaning, and your pairing that sentence off with the right meaning is obviously different from an image of that sentence flashing through your mind. If I hear a sentence of a language that I don’t understand, it may pass through my mind later (i.e., I’ll remember the sounds later, just as songs that one has heard often flash through one’s consciousness), but I won’t know what it means. So understanding a sentence is different from its flashing through your mind. But if your thinking *if snow is cold, then snow is not hot* were identical with your thinking “if snow is cold, then snow is not hot,” then your having that thought *would* be identical with that sentence’s flashing through your mind. Since it isn’t, we don’t think in natural language.

4.0 Why version #2 is false

Do we think in some innate code? No. (Some of the arguments that we used against Version #1 obviously apply to Version #2. So we won't repeat those.)

4.1 First reason

We often think in images (visual and auditory), and the contents of these images seem to have a structure that is fundamentally different from the structure of any sentence. Sentences are “digital” structures, meaning that they decompose into a finite number of distinct parts. Sense-perceptions are non-digital; they are “analogue” structures, meaning that they *don’t* decompose into discrete parts and are thus characterized by a kind of seamlessness that sharply distinguishes them from sentences.

4.2 Second reason

Knowing a language involves understanding the expressions belonging to it. Understanding such expressions involves knowing the rules that assign meaning to them. One can’t understand any language, including an innately known one, without knowing such rules. But if one’s thought is entirely dependent on such a language, then one can’t know those rules, or anything else for that matter, without understanding those expressions. Therefore it is viciously circular to hold that we think in an innately known code.

4.2.1 How advocates of the view in question deal with this argument

Advocates of the view that we think in an innate code deal with this problem by saying that, strictly speaking, we *don't* understand the sentences of our innate code. We are merely "built to conform" to those sentences. In other words, although we don't understand those sentences, we are so neurologically structured that they cause us to act in ways that are appropriate to the meanings of those sentences. So if the sentence of my innate code that means *fire is hot* goes off in my head that fact will *cause* me to go near fire if I'm cold and to stay away from it if I'm hot (or some such).

4.2.2 Why this move doesn’t work

If a sentence isn’t understood by so and so, then from so and so’s viewpoint it is not functioning *as* a sentence. If, as advocates of the viewpoint just presented maintain, nobody understands the sentences that mediate their thoughts, then those sentences aren’t really functioning *as* sentences and, for all intents and purposes, they cease to be sentences.

Here’s an analogy. Suppose that you have a book, but that, instead of reading it, you use it as a weapon—you throw it at your roommate, of whose antics you are growing increasingly tired. Even though the thing you are using as a weapon is a book, it is not in this context functioning *as* a book and, for all intents and purposes, is not a book. Similarly, if nobody understands the sentences that are, supposedly, identical with their own thoughts, then those sentences aren’t really functioning as sentences anymore—they’re functioning more like the copy of *War and Peace* that, instead of being read, is being hurled at one’s roommate.

Most importantly, the view that our thoughts are sentences that we don’t understand is deeply implausible, if not tautologically false.

So we may conclude that we *don’t* think in language, even though, undeniably, knowledge of language enormously enhances at least some of our cognitive capabilities.

5.0 Given that thoughts aren't identical with sentence-tokens, why is it that knowledge of language enhances cognitive ability and why does it feel as though we think in words?

On your computer desktop, there are various "icons." By manipulating these icons, you can perform various operations—you can close a document, open your email account, etc. But these icon-manipulations are but representations of changes in patterns of electrical activity to which they bear only an extremely schematic resemblance.

When I'm thinking—when, for example, I'm trying to solve a problem (especially one connected to my work)—my thoughts often take the form of a voice. I have mental images of spoken words. This also happens when I'm reading. Sometimes it happens that I read one book for a while (e.g., one by Frege) and then start reading another (e.g., one by comedian Dave Barry), and I'm startled as to why *Frege* would say such things. Then I realize that I'm no longer reading Frege and that I should therefore "switch voices."

And, of course, people think with the help of visual imagery. Many a student of calculus, or even of more advanced branches of mathematics, relies on mental images. I know I do.[206] But the images that go through my head when solving some problem aren't *identical* with the ratiocinative activities in which my problem-solving thought consists. Taken by themselves, those images are quite feeble. They don't specify the exact shape of the object I'm thinking about. I'll use a five-sided figure to stand for a ten-sided figure, since I can't visualize the latter. But it doesn't matter, provided that I can find *some* mental icon that will stand for the relevant concept. The icon does not itself have to depict the concept or object it represents. It need only be sufficiently differentiated that I won't confuse it with the image I am using to stand for some other concept of mine.

But even if I were more visually adept and, for that reason, could produce more exact visual surrogates for the concepts I'm thinking about, my ratiocinative activity couldn't possibly be *identical* with any sequence of images. For any given sequence of images, a creature that could experience those images wouldn't necessarily have any of the concepts that they represent in my thought processes. No matter how *conceptually* impoverished a creature is, little or nothing follows as to what sorts of images can go off in its head.

But everything constitutive of *consciousness* seems to be image-like or, at

the very least, to share with images the property of being ***phenomenologically*** pregnant.[207] Wittgenstein (1958) concluded from this that thinking isn't a private psychological act. For reasons that we'll discuss in the next section, he inferred that thinking consists of engaging in overt behaviors that involve symbols belonging to some public language.[208]

Wittgenstein's position is rank absurdity. He's right, of course, that tickles, itches, etc., aren't thoughts.[209] But what follows isn't that thoughts aren't mental entities. What follows is that thoughts, unlike itches and tickles, aren't phenomenologically pregnant.

Incidentally, the mistake Wittgenstein is making here is very similar to a mistake that David Hume makes in connection with the nature of personal identity. David Hume argues that there is nothing to one's mind other than the conscious events that populate it.[210] His argument is that, when he introspects, he only encounters various tickles, pains, and other sensations. This argument assumes that everything there is to know about one's mind is to be known in the same quasi-perceptual, phenomenology-drenched way in which we know of our sensations. But this assumption is blatantly false. I know that I believe that 2 + 2 = 4, and I know it in a relatively direct way. But it isn't because I experience some *sensation* or mental image or other phenomenologically pregnant entity that I know this.

Wittgenstein's view that thinking consists of manipulating expressions is inconsistent with the fact that, unless one can think, one cannot use expressions meaningfully. If I just bark out sounds, having no idea what they mean, I'm not in any real sense using language. For my noise-making to be speech, it must be guided by a knowledge of what those noises mean.

Wittgenstein is right that the sensations and images that pass through one's mind when one is having a given thought do not *constitute* that thought. But he's wrong to conclude that thinking isn't a psychological process. (As previously stated, what he should have concluded is that not all psychological activity consists of images and sensations.) He's also wrong to conclude that those images categorically have no significant cognitive function. Mental imagery facilitates thought. Thanks to it, we have thoughts that we otherwise simply couldn't have.

Given this data, I would suggest the following. Though not themselves constitutive of thought, mental images allow us to manage thought-processes that we otherwise couldn't manage. Just as one can manipulate otherwise uncontrollable patterns of electrical activity by manipulating the icons on

one’s computer desktop, so one can manipulate otherwise uncontrollable patterns of ratiocinative activity by manipulating cognitive imagery.

There is another relevant point. Consciousness is a place where there are different streams of cognitive activity—where otherwise mutually isolated streams of cognitive activity can meet. The processes that create an olfactory or auditory perception are isolated from those involved in reading music or understanding facial expressions. But, in consciousness, sense-perceptions meet with judgments and with intentions to act, and so on. Consciousness has the effect, and maybe the purpose, of integrating otherwise mutually isolated streams of cognitive activity and, therefore, of pooling otherwise discrete bodies of knowledge. Mental imagery improves thought because it enables otherwise compartmentalized ratiocinative activities to borrow information from other similar activities from which they would otherwise be sealed off.

By virtue of knowing a language, one automatically has at one’s disposal a rich and easily operated system of icons by means of which one can manage one’s cognitive activities in the way just described. This, I would suggest, is why knowledge of a language does so much for thought.

6.0 Wittgenstein on language, thought, and the relationship between the two

Ludwig Wittgenstein (1889–1951) put forth several provocative contentions about language.[211] Perhaps the most celebrated of these is his contention that there cannot possibly exist such a thing as a *private language.* Human beings can devise private codes, he granted. But, Wittgenstein says, any such code must be a translation of an existing public language. It isn't even theoretically possible, Wittgenstein thought, for there to exist a language that wasn't created by several interacting individuals.

Once such a language is created, Wittgenstein grants, a single person can use it on his own—he can go off into the woods and keep records with it. But any case of a single person's using a language for his own private use is, in Wittgenstein's view, parasitic on there being, or once having been, a multiplicity of people using that same language.

Wittgenstein isn't making the pedestrian point that, as a matter of psychological fact, it would be hard, maybe even impossible, for somebody to single-handedly create a language (other than one that was a translation of an existing language). He's saying that it's logically impossible—that such a language is a ***surd,*** like a square circle.

Wittgenstein also held that one can't think unless one knows a language. Wittgenstein seemed to grant that one could have mental states without knowing a language. But he pretty clearly thought that knowledge of a language was a prerequisite for any sort of reasoning.

If Wittgenstein is right, one cannot think unless one is embedded in a society of some kind. It is not just psychologically, but also logically impossible, Wittgenstein thought, for a human being who isn't either a current or erstwhile member of a society to think. These contentions, along with Wittgenstein's arguments for them, are put forth in his book *The Philosophical Investigations.*

Wittgenstein provides an interesting argument for his view that there cannot be a private language, which is known, appropriately enough, as the Private Language Argument (PLA). We'll discuss PLA in a moment.

Wittgenstein provides two distinct arguments for the position that one cannot think without knowing a language. Because of the similarities between them, contemporary authors tend not to distinguish these arguments from one another, and they are collectively referred as to the Rule Following

Argument.

6.1 Wittgenstein's first Rule-Following Argument

> (RF[212]) Not all ideation is ratiocinative. For ideation to be ratiocinative, it must be rule-governed. That is, it must be guided by a knowledge of the canons of logic. Thus, thinking involves following rules.
>
> But there is no psychological condition or event that necessarily accompanies an act of following a rule. To see this, go ahead and follow some rule. (Add two multi-digit numbers together; read some passage out loud.) When you introspect, you may find that you are experiencing these or those feelings or images. But you could obviously follow the rule in question without having such experiences. Thus, following a rule is not a psychological act.

Contrary to what this argument assumes, not all mental states are images or sensations.[213] Your belief that 1 + 1 = 2 isn't such a thing; and your deploying that belief on some specific occasion isn't identical with your experiencing some image or sensation or series of images or sensations.

Not everything in your mind is a fleeting conscious event. Much of what constitutes your mind consists of stable, enduring structures. Your belief that snow is white isn't a flash in the pan. It's been there for a while, and will continue to be there for a while and though it gives rise to momentarily conscious events, it isn't identical with any such event or with any aggregate of such events. Contrary to what Wittgenstein alleges, given only that there is no tickle or itch or other fleeting mental event that necessarily accompanies an act of rule-following, it doesn't follow that rule following isn't a psychological act

6.1.1 Wittgenstein's second Rule-Following Argument

(RF_2) Consider the sentence "rabid dogs always foam at the mouth." For argument's sake, let's suppose that there exists some entity *M* that is the meaning of this sentence. Let's also suppose that, without grasping or operating with the sentence "rabid dogs foam at the mouth" or any other expression, you behold *M*—you *grasp* it. You *see* it—through your "mind's eye," or some such. Finally, let's suppose that you not only grasp *M*, but know it to be true. In other words, you know that rabid dogs foam at the mouth.

If you can't see that, given *M*, Fido can't be rabid without foaming at the mouth, your grasp of *M* is, to that extent, useless and as good as non-existent. Thus, so far as you don't grasp how *M* bears on other meanings—so far as you don't grasp its *significance*, in other words—your grasp of *M* is useless and might as well not exist. The problem we had with the sentence "rabid dogs foam at the mouth" arises in connection with *M*. Wanting to explain what it is to understand that sentence, we posited *M*. But it turned out that *M* is just like that sentence: it's useless unless *its* meaning is grasped. *M* turned out to be just another symbol—just another cognitively dead symbolic intermediary.

Meanings are themselves symbols. So unless we already know what it is to understand symbols, positing symbol-meanings doesn't help us understand the nature of symbol-comprehension. Positing symbol meanings in the hopes of explaining symbol-comprehension is like positing little people inside people's heads in the hopes of explaining how people think.[214] Positing these little people is useless unless it's *already* understood how people think. Positing meanings is useless unless it's *already* known how symbols are understood.

6.1.2 Wittgenstein's second rule-Following agument criticized from a Platonic viewpoint

This argument of Wittgenstein's bears a striking resemblance to Aristotle's Third Man Argument (TMA).[215] The purpose of TMA is to undermine Plato's contention that spatiotemporal entities are instances of properties and, in addition, that the latter are not spatiotemporal. It will help if we take a moment to consider Aristotle's argument (it will soon be clear why the third and fourth sentences are italicized):

> (TMA) For argument's sake, suppose that there is such a thing as the property of being a triangle. Let T_1 be that thing. If Platonism is right, *T_1 must itself be a triangle. Further, if Platonism is right, any case of two objects' sharing some property is a case of there being some third object of which they're both instances.* Thus, if T_2 is some triangle, there must be some third object T_3 such that each of T_1 and T_2 is an instance of T_3. But, by the same logic, there must be some fourth thing T_4 such that each of T_1, T_2, and T_3 is an instance of T_4. And so on *ad infinitum.*

Aristotle's argument assumes that properties are instances of themselves—that the property of being a triangle is itself triangular. But that's precisely what Platonism denies.[216] Platonism is the doctrine that spatiotemporal things are but *instances* of non-spatiotemporal things, the latter being properties. The property of being spatiotemporal is therefore itself non-spatiotemporal and, therefore isn't an instance of itself. And the property of being a cat obviously isn't a cat, since cats are spatiotemporal, whereas that property is not.[217]

Just as, according to Aristotle, Platonism is struck with the problem of explaining why the property of being a triangle was itself a triangle, so according to Wittgenstein, those who believe in meanings are stuck with the problem of explaining how meanings are understood.

But meanings are not themselves understood. They're no more understood than the property of being a cat is a cat. Consider the expression "I." I'm referring, not to specific occurrences (utterances, inscriptions) of that

expression, but to that expression itself—the ***expression-type,*** in other words, as opposed to its ***tokens.*** The meaning of that expression is some rule that assigns referents to its tokens depending on the circumstances. More precisely, the meaning of "I" is a rule to the effect that a token *t* of that expression that is uttered by *x* refers to *x*. My knowing that "I" has that meaning is part of what enables me to understand utterances of "I'm getting really bored," "I really like snow-boarding," etc. But that meaning is not itself understood. Meanings are either grasped or they're not. There's no such thing as grasping a meaning but not understanding it.

A consequence is that meanings don't have to be *interpreted*. It's incoherent to suppose that Hoigaard (who, let's assume, is Norwegian) should be able to grasp this meaning, but fail to see that people were referring to themselves when they said "I."

This isn't to say, what would clearly be false, that one automatically recognizes every consequence of every principle that one grasps. People fail to see the consequences of the principles they grasp. In fact, since any given proposition has infinitely many consequences, people fail to grasp infinitely many consequences of anything that they grasp. But *to the extent* that one's grasp of a given principle is determinative of what one knows, there isn't a gap between grasping meaning, on the one hand, and applying it, on the other. And to the extent that one doesn't see the consequences of some meaning or principle that one grasps, it isn't because one has failed to *interpret it.*

A story will clarify these points. Brown is a highly (but not superhumanly) intelligent person who not only speaks perfect English, but is exceptionally (but, again, not superhumanly) adept at understanding highly intricate sentences. *XYZ* is a sentence of English that is 100,000 words long but is otherwise normal. *XYZ* is grammatical, coherent, etc.[218] Even though Brown knows the relevant semantic rules, he obviously can't understand *XYZ*. But the problem isn't that Brown can't interpret the relevant semantic principles. The problem is that, because that sentence is so long, he can't *deploy* his semantic knowledge. In and of itself, that semantic knowledge is good to go. But, under the circumstances, it's weighed down by limitations on Brown's part that have nothing at all to do with his *semantic* competence. Brown's memory isn't perfect; he can't keep too many tasks going in his mind simultaneously, even if any given one of them is elementary. These aren't *semantic* limitations on Brown's part. The problem isn't that he doesn't know

the relevant semantic rules. Nor is it that he hasn't adequately *interpreted* the semantic rules that he grasps. The problem is that his grasp of those principles is being inhibited by shortcomings that in no way detract from his semantic competence, even though they *do* limit his ability to *deploy* it. *To the extent that* what Brown knows is determined by his grasping those semantic rules, Brown *does* understand the expressions to which they assign meaning. But, of course, what Brown knows is a function of many things besides his grasp of those principles.

Wittgenstein's idea that meanings, if grasped, must be *interpreted* reflects his erroneous belief that, whenever one can't see a consequence of some principle that one grasps, it's because one hasn't adequately interpreted that principle. But principles aren't interpreted. They're grasped. And, if grasped, they guide judgments. And to the extent that one's grasp of a principle is determinative of what one knows, those judgments are consistent with those principles. But, of course, one's grasp of a given principle is but one factor among many determining the course of one's thought-process; so one's grasp of that principle may be prevented from eventuating in the judgments in which it would otherwise eventuate. But when that happens, it isn't because that principle wasn't adequately *interpreted.*

6.2 The private language argument

Neither of Wittgenstein's attempts to show that thinking isn't a psychological act is probative. They're interesting arguments, if I've interpreted them correctly; and if I haven't, it's an interesting question how they ought to be interpreted. But given what we've seen, we have no reason to reject the plausible, if not self-evident, position that thinking is a psychological act.

But, of course, Wittgenstein doesn't see it this way. Believing himself to have definitively established (the radically absurd principle) that rule-following and, more generally, thought aren't mental processes; he attempts to show that there can be no private languages. Here is his argument:

> (WA)[219] Imagine the following. Smith is alone on a desert island. He's always been this way and, therefore, has never learned a language. But he wishes to create one. He thus decides to stipulate that the meanings of expressions $X_1...X_n$ are to be respectively, *tree, berry, sand, sun,* etc. (He also specifies how those symbols are to be written down. For some reason, he has a notebook and a pen.) He does this all on Monday. We'll refer to the language Smith creates as "Smithese$_1$." One of the semantic rules for Smithese$_1$ is to the effect that:
>
> X_{35}: means that the berries on the dark side of the hill are poisonous.
>
> That same day, Smith sees some poisonous berries on the dark side of the hill, and writes X_{35} down in his notebook. (Let X_{35} be our symbol for that particular inscription.) On Tuesday, he reads what he wrote down. Question: is there any way that Smith can misinterpret what he's written? If he takes X_{35} to mean that the sun is out, his interpretation of X_{35} will be inconsistent with his Monday intentions. But so what? Smith is the only person who speaks Smithese$_1$, so the expressions belonging to Smithese$_1$ mean whatever Smith thinks they

mean.

It follows that there could not exist a language that, like Smithese$_1$, is private. An English-speaker can't just decide that "snow is white" means that grass is green. For it is the thoughts and practices of people in general, and not of this or that specific person, that determines what is meant by "snow is white." But in this context, Smith *is* "people in general." So his believing that X_{35} means that the sun is out is like *everybody's* deciding that "snow is white" means that grass is green. And just as "snow is white" *would* mean that if everybody thought it did, so X_{35} means that the sun is out because, in effect, everybody thinks it does, the reason being that, in this context, Smith *is* everybody.

Thus, the expressions of Smithese$_1$ mean whatever Smith thinks they mean. This means that he can't be wrong as to what they mean.

But a symbol that can't be used wrongly doesn't mean anything. What makes utterances of "John is a soccer player" meaningful is that they are used wrongly if John isn't a soccer player. If "John is a soccer player" couldn't be used wrongly, you couldn't conclude anything from the fact that such an utterance was correct—you would be no less entitled to conclude that the moon was made of Styrofoam than you would be to conclude that the dog ate your wig.

Thus, there are languages only where there are right and wrong ways to use expressions. Since it is only where several people are using a language that there are right and wrong ways of using its expressions, it is only where there are several people that there are languages: all languages are public.

WA is spurious and establishes nothing. First of all, unless we can trust our sense-perceptions and our memories, the testimony of other people as to what words mean is worthless. If I can't trust the sense-perceptions involved in my being told what the word "grape" means, and if I can't trust the memories deposited by those perceptions, then I have no reason to give any credence to the testimony of the people who defined that word for me. So the

supervision of others is useful *only to the extent* that I can trust my memories. But to the extent that I can trust my memories, I *am* competent to remember how I've defined words belonging to a private language that I've invented; and Wittgenstein's argument implodes.

But Wittgenstein's argument involves other, equally profound blunders. The reason that expressions of English (or any other public language) are meaningful is that people remember what they mean. If everybody forgot what "snow" meant, it wouldn't be of much use. To be sure, where English is concerned, no one person's forgetting the meaning of a word is going to strip of it meaningfulness, since there are millions of others who still know it. But that only means that English is causally more deeply entrenched than a language, such as Smithese$_1$, that only one person speaks. This is not to mention that, if *I* forget the semantic rules for English, it will be as useless to me as Smithese$_1$ would be to Smith if he forgets the semantic rules of Smithese$_1$. By the same token, if Smith remembers the semantic rules for Smithese$_1$, the expressions of that language *will* be useful for him. He'll know not to eat the berries on the dark side of the hill; he'll know where the good-tasting aardvarks are; etc.[220]

What about Wittgenstein's point that the expressions of Smithese$_1$ mean only what Smith thinks they mean? A language *is* a set of semantic rules. If, on Monday, X_{35} means that the berries on the dark side of the hill are poisonous, whereas on Tuesday that same inscription means that the sun is out, then the language Smith is using on Tuesday is *ipso facto* a distinct one from the language he is using on Monday. On Tuesday, he's speaking Smithese$_2$, not Smithese$_1$. Therefore, X_{35} means that the sun is out *only* if it isn't an expression of Smithese$_1$. The expressions of Smithese$_1$, then, don't mean whatever Smith thinks they mean. If his opinions assign new meanings to X_1...X_n, those opinions are creating a new language that happens to be orthographically and phonetically coincident with Smithese$_1$; those opinions aren't assigning new meanings to expressions constituting some existing language. So Wittgenstein is just wrong to say that the expressions of Smithese$_1$ mean whatever Smith thinks they do.[221]

Thinking seems to involve following rules. Wittgenstein grants that people think. But since he believes totally absurdly, that rule-following and, therefore, thinking aren't psychological acts, he is forced to come up with a

different analysis of what thought is. He wrongly says that thought consists in the manipulations of symbols. Given his belief that there aren't private languages, he holds that to think is to manipulate symbols of public languages.

Given that RF and WA are spurious arguments for false conclusions, Wittgenstein has given us no good reason to accept this position. In any case, we know it to be an erroneous one. A prerequisite for speaking a language is knowing its semantic rules and being able to operate in an intelligent and, consequently, thought-mediated fashion.

Chapter 9 Modality and Non-Existence

1.0 Necessity and possibility properties of statements, not objects

A modal statement is one that says what *could* be the case or what *must* be the case. Here are some examples of such statements:

(1) Triangles must have three sides.
(2) It's possible to make money without selling out.
(3) There's no way anyone could have survived that fall.
(4) Assuming that Bob has fifty cars, he must have at least some money.
(5) John's becoming stronger was an inevitable consequence of his lifting weights all summer.

For any proposition *P*, it's necessarily the case that *P* ***iff*** it's not possible that not-*P*; and it's possible that *P* iff it's not necessarily the case that not-*P*. ("Iff" is short for "if and only if.") It's necessarily the case that Jim is in the barn iff it's not possible that Jim is not in the barn; and it's possibly the case that Jim is in the barn iff it's not necessarily the case that Jim is not in the barn.

There are no impossible objects (no objects that couldn't possibly exist). And there are no possible but non-actual objects (no objects that could exist but don't). Possibility and impossibility are properties of propositions, not of objects. For example:

(6) "square circles cannot possibly exist,"

says that the proposition:

(7) there are square circles,

cannot possibly be correct. (6) doesn't make the self-contradictory claim that there exists some object *x* such that *x* is a square circle and such that *x* cannot exist. And:

(8) "happiness is a possibility for Jim, even though he's still miserable,"

says that, even though Jim is now miserable, the proposition:

(9) Jim is happy,

could be true. (8) doesn't make the absurd claim that there exists some *x* such that *x* is Jim's as of yet nonexistent happiness and such that *x* could exist.

1.1 Existence and non-existence properties of propositions, not of objects

The points just made concerning the concepts of possibility and impossibility are to be understood in terms of more general points concerning the concepts of existence and non-existence.

Nothing doesn't exist. There exists no *x* such that *x* doesn't exist. Statements that, given their surface-structures, appear to attribute non-existence to objects in fact attribute falsity to propositions. A story will make this clear.

An unscrupulous person claims to have been attacked. In reality, he wasn't attacked; but, for whatever reason, it's to his advantage for people to believe otherwise. He describes his non-existent assailant to a police sketch artist. The resulting sketch is distributed across the nation. If we say that the sketch depicts a nonexistent person, we're saying that there exists some person who doesn't exist who it depicts. But that would be absurd. The sketch is better thought of as saying: *there is somebody having such and such features,* these being the features that a person *would* have to have *if* he were to fit the sketch. Though false, that *statement*—that ***existence-claim--***obviously exists.

Similarly, the cartoons that brought Fred Flintstone and Bart Simpson into our homes can be thought of as making false statements of a similar kind. In the case of Fred Flintstone, that statement is along the lines of: *there exists a boorish man who has a pet dinosaur named "Dino," a wife named "Wilma,"* etc. In the case of Bart Simpson that claim makes a statement along the lines of: *there exists a bratty boy who torments his oafish father with his loveable shenanigans,* etc.

All of these *statements* exist; they're simply not true. Nothing answers the descriptions encoded in them. Therein lies the real meaning of "Fred Flintstone/Bart Simpson doesn't exist." The real meaning is clearly not some absurd claim to the effect that there exists some entity that doesn't exist.

2.0 Necessity does not come in degrees

Here is a view about modality that is false, as we'll see in a moment, but is nonetheless a useful starting point[222]:

> (DN) There are different degrees of necessity. "Squares have four sides." This is necessary in the strongest sense. The very idea of a square's failing to have four sides is an incoherent one.
>
> "Metal expands when heated." This is not *as* necessary as "squares have four sides," since the idea of a universe in which metal didn't expand when heated is not incoherent. But, given that it describes a law of physics, and is therefore utterly unalterable, it clearly has a high degree of necessity.
>
> "After surgery, people are unusually prone to get infections and must be given antibiotics to stave them off." There is clearly *some* sense in which this is necessarily true. An immunologically normal person who *wasn't* more prone to get infections after surgery is a medical impossibility. And, unless he qualified his point very heavily, a professor of medicine who denied this would be misleading his students. But this statement clearly isn't *as* necessary as the others.

DN is false. There are only two degrees of necessity: having it completely and lacking it completely. When we say that *P* is "more necessary" than *Q*, we're saying something about the conditions that must hold for those statements to be true; we're saying that the conditions that must hold for the first to be true are weaker than those that must hold if the second is to be true. "Squares have four sides" is maximally necessary, because it holds *unconditionally.* In any universe at all, no matter how mass-energy is distributed in it, that statement is true.[223] So *no* condition has to be met for that statement to be true. Or, to think of it another way, the only conditions that have to be met are maximally weak: they could be thought of as limiting cases of conditions that are trivially satisfied by any distribution of mass-energy. To say that "metal expands when heated" is "less necessary" than "squares have four sides" is to say that the former could to be true in a wider class of circumstances than the latter; and to say that "after surgery, people are unusually prone to get infections" is "less

necessary” than “metal expands when heated” is to say that the former could fail in a wider class of circumstances than the latter.

So there aren’t really degrees of necessity. Something is either necessary or it isn’t. But some statements (e.g., “two is bigger than one”) need less help from others to be true—those are the “highly necessary” ones. Others (e.g., “people who wear hats are bad drivers”) need a little more help—those are the “less necessary” ones.

3.0 No necessary *a posteriori*

In 1969, Saul Kripke gave a now famous series of lectures in which he argued that there are non-analytic necessary truths. Here, in my words, is Kripke's argument ("KA" for short):

> (KA) The ancients referred to the last celestial body (besides the Sun) to disappear from the morning sky as "Phosphorous" (or, to be precise, they referred to it with an expression whose Anglicization is "Phosphorous"), and they referred to the first celestial body to appear in the evening sky as "Hesperus" (same qualification). The former is Venus and so is the latter, and the former is therefore *identical* with the latter. The ancients didn't yet know this; and they went for a very long time having no idea that some *one* celestial body appeared in the evening sky before any other, and disappeared from the morning sky after any other. But eventually they found this out. What they found out isn't trivial, of course; it isn't in the same category as the fact that Hesperus is identical with Hesperus. What they discovered is a non-trivial, empirical fact.
>
> Henceforth, "FES" and "LMS" will be our abbreviations for the predicates "first celestial body to appear in the evening sky" and, respectively, "last celestial body (besides the Sun) to disappear from the morning sky."
>
> Contrary to what it is extremely tempting to believe:
>
> (HIP)[224] "Hesperus is Phosphorous,"
>
> does *not* have the same meaning as:
>
> (S*) "the LMS is identical with the FES."
>
> It will take a moment to make it clear why this is so. Let *P* be the proposition expressed by HIP. As we just saw, empirical investigation was needed to establish the truth of *P*; so *P* is an empirical truth and is therefore non-analytic. It's important to

keep the distinction between HIP and *P* in mind. The ancients didn't have any feelings about the English sentence "Hesperus is identical with Hesperus." And, in making the astronomical discovery that we've been discussing, what they found out had nothing to do with HIP—or any other sentence. It was an astronomical, not a linguistic, discovery. This point is crucial.

Moving on: If HIP has the same meaning as *S**, then *P* is identical with the proposition meant by *S**. And the latter is equivalent with the proposition that:

(*P**) exactly *one* thing is an LMS and exactly one thing is an FES, and nothing that has the one property lacks the other. [225]

The same is true of *P*, supposing that *S** and HIP express the same proposition.

In light of these points, consider a world W* satisfying the following conditions. *W** is *semantically* just like our world. (Two worlds or situations are "semantically identical" if the rules that govern word-usage in the one coincide with those that govern it in the other.) So, in *W**, people speak English. But, in *W**, *Venus* isn't the LMS or the FES. In *W**, because of some cataclysm that didn't occur in our world, Mars is the LMS, and Mercury is the FES. So, in *W**, "the LMS" picks out Mars, and "the FES" picks out Mercury. This follows from the fact that *W** is semantically just like our world.(If John McCain had become President in 2008 instead of Barack Obama, "the U.S. President" would, given the semantic rules that we use, refer to McCain. But, given those very same rules, "Barack Obama" would still refer to Barack Obama. In other words, if *x* is the person who *in our world* is referred to as "Barack Obama," "Barack Obama" would refer to *x* in *W**. Similarly, if Mars were the LMS, as opposed to Venus, "the LMS" would, given the semantic rules that we use, refer to Mars. But given those same rules, "Venus" would still refer to Venus. In other words, if *x* is the planet that *in our world* is referred to as "Venus," "Venus" would refer to *x* in *W**.

“Venus,” like “Barack Obama,” is thus a *rigid designator*. If *x* is what “Venus” *in fact* refers to, *x* is what “Venus” would refer to in *any* world semantically like ours. But “the LMS,” like “the current U.S. President,” is *not* a rigid designator; it refers to different things in semantically identical situations. (It used to refer to Carter; now it refers to Obama. But the semantic rules governing that expression haven’t changed in the least in the meantime.)

In general, proper names are rigid designators. “Hesperus” and “Phosphorous” are proper names. Therefore, *unlike* “the LMS” and “the FES,” they are rigid designators. So as long as the semantic rules of English haven’t changed, there is some one object *x* such that:

(i) “Hesperus” refers to *x*, and so does “Phosphorous”;

and, consequently, such that:

(ii) each of ‹Hesperus has phi› and ‹Phosphorous has phi› is true exactly if *x* has phi;

a corollary being that:

(iii) “Hesperus is Phosphorous.”

It is obvious that, for any object *x*, the proposition *x is identical with x* is necessarily true. But it’s equally obvious that the proposition expressed by HIP is empirical, and therefore non-analytic. Therefore, HIP expresses a non-analytic *necessary* truth.

3.1 Kripke's argument evaluated

This argument involves a massive non-sequitur. A story will make this clear. I meet a man. We shake hands. I obviously see what he looks like. He tells me that his name is "Hesperus." (Without worrying about the details, let's assume that I know him to be telling me the truth.) In this context, there is some individual *x* such that I am learning that:

(HX) "Hesperus" is *x*'s name.

But *what* I am learning must be distinguished from the information *through which I am learning it.* The information relayed to me by my senses isn't confined to HX. It would be absurd to think otherwise. That information is more along the lines of:

> (BHX) There is some individual *x* such that *x* is standing right in front of me (in my office in Richmond, VA) at this moment (3:00 P.M., March 13, 2009) and such that *x* has such and such characteristics (he has a mustache, he's over six feet tall, etc.) and such that I am being told, by *x* himself, that "Hesperus" is *x*'s name.

So there is some *x* such that "Hesperus" merely labels *x*. "Hesperus" doesn't have for any part of its meaning anything having to do with the property of being over six feet, having a mustache, being in this or that place at this or that time. But the information *through which* I learn whom "Hesperus" labels *does* concern such properties. That information is descriptive; a certain person *x* is visually represented, and thus described, to me as having certain characteristics (over six-feet tall, etc.); and, with the help of this information, it is also being made clear to me that "Hesperus" names *x*. But that descriptive information is not *itself* a part of the semantic rule that, with its help, I am learning—that is to say, it is not itself any part of HX. HX is extremely threadbare; it merely puts a label on a certain object. So the just-mentioned descriptive information has no place in the *semantics* of "Hesperus"; its relevance is confined to the pre-semantic act of assigning it to the right object. That information isn't semantic; it's *pre*-semantic.

The following evening, I receive a phone call. I don't recognize the

caller's voice. The caller is obviously using a voice-modulation device, giving it a metallic, threatening sound. And threaten is just what he does; for he says "if you don't move out of town, I'm going to kill you. By the way, my name is Phosphorous." I look at my caller ID; and, sure enough, the call is from one "Phosphorous."

In this context, there is some individual *y* such that I am learning that:

(PX) "Phosphorous" is *y*'s name.

PX merely registers the fact that a certain name is associated with a certain object. It doesn't say anything (else) *about* that object. It doesn't say whether that object is tall or short, friendly or unfriendly. PX is exceedingly threadbare. But the information *through* which I am learning PX is very different from PX itself. *That* information—that *pre*-semantic information—is along the lines of:

(BPX) There is some individual *y* such that, a moment ago (7:00 P.M., March 14, 2009), *y* called me; such that, using a voice-modulation device to disguise his voice, *y* is making threatening statements; and such that I am being told, by *y* himself, that "Phosphorous" is *y*'s name.

Here must we repeat (*mutatis mutandis*) what we said a moment ago. There is some *y* such that "Phosphorous" merely labels *y*. "Hesperus" doesn't have for any part of its meaning anything having to do with the property of talking on the phone with me at a certain time, or of using a voice-modulation device, or of making threats. But the information *through which* I learn whom "Phosphorous" labels *does* concern such properties. That information is descriptive; a certain person *y* is perceptually represented, and thus described, to me as having certain characteristics (calling me at a certain time, making certain statements, etc.) and, with the help of this information, it is also being made clear to me that "Phosphorous" names *x*. But that descriptive information is not *itself* a part of the semantic rule that, with its help, I am learning—that is to say, it is not itself any part of PX. PX is extremely threadbare; it merely puts a label on a certain object. So the just-mentioned descriptive information has no place in the *semantics* of "Phosphorous," its relevance is confined to the *pre*-semantic act of assigning

it to the right object. That information is entirely *pre*-semantic.

But such *pre*-semantic information can have an incalculably profound effect on what is (non-semantically; i.e., non-literally) *conveyed* by utterances. A continuation of our story makes this clear. Right after getting the upsetting phone call, the police call. For some reason, they know about the phone call. They say to me:

(A_1) "Phosphorous is a professor of political science."

There is some *y* such that A_1 does nothing more than affirm the innocuous proposition that:

(A_2) *y* is a professor of political science.

But under the circumstances, what will be conveyed to me is some much richer proposition along the lines of:

> (A_3) There is some individual *y* such that, a little while ago (at around 7:00 P.M., March 14, 2009), *y* rang me; such that, using a voice-modulation device to disguise his voice, *y* is making threatening statements; such that I am being told, by *y* himself, that "Phosphorous" is *y*'s name; and, finally, such that *y* is a professor of political science.

The pre-semantic information embedded in BPX affects what I *take away* from the policeman's utterance of A_1, even though it is no part of its literal meaning. For exactly similar reasons, the pre-semantic information embedded in BHX affects what I would *take away* from an utterance of:

(B_1) "Hesperus is an avid tennis player."

There is some *x* such that B_1 does nothing more than affirm the descriptively impoverished proposition that:

(B_2) *x* is an avid tennis player.

But under the circumstances, what that utterance conveys to me is some much richer proposition along the lines of:

> (B_3) There is some individual *x* such that, a couple of days ago (March 13, 2009) at about 3:00 P.M., *x* is standing right in front of me (in my office in Richmond, VA); such that *x* has such and such characteristics (he has a mustache, he's over six-feet tall, etc.); such that "Hesperus" is *x*'s name; and such that *x* is an avid tennis player.

We need only add one finishing touch to our story to see what's wrong with Kripke's reasoning. The police arrest Phosphorous. They ask me to come down to the station to talk to them. They point to Phosphorous (who is handcuffed) and say: "that's Phosphorous." Phosphorous, I now see, is none other than Hesperus. The man I see is the mustachioed man I met a few days earlier. The police then say: "Phosphorous also goes by the name of Hesperus. His legal name is 'Hesperus Phosphorous.' But he tends to introduce himself as either 'Hesperus' or 'Phosphorous.' In any case, *Hesperus is Phosphorous*." The italicized sentence is none other than HIP, of course.

Given what we've already learned, it's obvious that there is some *one* individual *z* such that this utterance of HIP is true exactly if:

> (K) *z* is identical with *z*.

Any proposition of that form is trivial. It is always trivial to say of an object that it is identical with itself. But by obvious extensions of what we said a moment ago, when I hear the policeman's utterance of HIP, what I *take* away from it—what it (non-literally) *conveys* to me, given the pre-semantic information at my disposal—won't be at all trivial, as it will be along the lines of:

> (K_2) There is some individual *x* such that, a couple of days ago (March 13, 2009) at about 3:00 P.M., *x* is standing right in front of me (in my office in Richmond, VA; such that *x* has such and such characteristics (he has a mustache, he's over

> sixfeet tall, etc.); such that "Hesperus" is *x*'s name; and such that there is some individual *y* such that, a few hours ago (at around 7:00 P.M., March 14, 2009), *y* rang me; such that, using a voice-modulation device to disguise his voice, *y* is making threatening statements; such that I was told by *y* himself that "Phosphorous" is *y*'s name; and such that *y* is a professor of political science; moreover, I have just learned from the police that *x is y*.

K_2 is an empirical proposition; it isn't analytic; a fortiori it isn't trivial. *But it's also a contingent truth.* K, on the other hand, *is* a necessary truth. *But it's also analytic.* So we don't have *anything* that is both necessary *and* non-analytic.

Kripke and many others have professed to find many examples of non-analytic necessary truths; but in each case, an exact analogue of the argument just given shows a confusion similar to the one just identified to be at work. In each case, semantics is being confused with pre-semantics. A defense of this claim is provided in Chapter 8 of my (2009) book *Analytic Philosophy* and in the first three chapters of my (2007) book *Conceptual Atomism and the Computational Theory of Mind*.

4.0 Propositional necessity as microstructural truth: Some preliminary concepts

The main modal concepts are necessity, possibility, impossibility, and contingency. All of these can be understood in terms of the concept of necessity. *P* is possible iff not-*P* isn't necessary. *P* is impossible iff not-*P* is necessary. *P* is contingent iff neither *P* nor not-*P* is necessary.

Whether all modal concepts can be understood in terms of this one concept depends on how inclusive one's conception of modality is. Like Bas van Fraassen (1989), I believe that the property of being probably true is a modal property—that, indeed, for any number *n*, the property of being probable to degree *n* is a modal property. It's an open question whether the concept of probable truth can be understood in terms of that necessary truth. I'm inclined to think that it can be, given how natural it is to think of necessity as a limiting case of probable truth. But until this is resolved, it can't be said that *all* modal notions can be understood in terms of the concept of probability. But there's no denying that necessity is the most important modal concept.

But what *is* necessity? One plausible answer is this: necessary truth is analytic truth; necessity is analyticity. "Squares have four sides" is "analytic," or "analytically true." Indeed, *all* analytically true statements are necessarily true. So it's obviously worth looking into the idea that necessity *is* analyticity. Many deny this (for reasons that we've already discussed); but many, including the present author, hold it. But different authors use the word "analytic" to mean different things, and we must make it clear how we will be using it before we can evaluate the proposal that necessary truth is analytic truth.

Sometimes that word denotes a property of sentences; other times, it denotes a property of propositions. A *sentence* is analytic if, given only what it means, it must be true. So "squares have four sides" is analytic. Given *only* what it means, it couldn't possibly be false. A *proposition* is analytic if it's non-empirically true. So the *proposition* that squares have four sides is analytic. No conceivable state of affairs could make it false. We will use the word "analytic" to denote a property of *propositions,* not sentences.

When examining the concept of analyticity, one must make it clear at the outset whether one is talking about sentences or propositions. Otherwise one won't be able to produce a coherent analysis of that concept. This is because there are analytically true *sentences* (or, strictly speaking, sentence-tokens) that have contingent propositions for their meanings. If, at this very moment, I were

to say “I am here now,” my utterance couldn’t fail to be true. Given the semantic rules that assign truth-conditions to that utterance, it’s a logical impossibility that it should be false. But *what it is that I’m saying* could very well be false. Given that I am JMK, that I’m in Richmond, and that it’s now March 13, 2009, I am, in producing such an utterance, attributing to JMK the property of being in Richmond, VA, on March 13, 2009. And, quite obviously, it’s *not* a necessary truth that I have that property. Similarly, if I say “I am not here now,” my utterance is analytically false. Given its semantics—given what the semantic rules are that assign truths-conditions to it—it’s logically impossible that it should be true. But what it is saying *could* easily have been false; I might not, after all, have been in Richmond, VA, at this very time. It isn’t in virtue of what it’s saying that a given token of “I am here now” (“I am not here now”) must be true (false); it is in virtue of *how it is saying it*. But it *is* in virtue of what it is saying that a token of “any case of knowledge is a case of true belief” must be true. This shows that there are two very different kinds of sentence-analyticity.

It also shows that sentence-analyticity diverges from proposition-analyticity and, more generally, that sentence-logic diverges from proposition-logic. In other words, the logical relations that propositions bear to one another don’t always parallel the logical relation that the corresponding sentences bear to one another. The two sets of interrelations can diverge quite dramatically. (Credit for this astonishing insight goes entirely to P.F. Strawson.) So it’s important to bear in mind that, in what follows, the word “analytic” refers to a property of propositions, not sentences. (Later we’ll discuss how it is exactly that analytically true (false) sentence-tokens can have contingent truths (falsehoods) for their meanings.)

The word “analytic” is ambiguous in another, equally important respect. As some philosophers use it, “analytic” is synonymous with “trivial” or “tautologous.” So “there are three feet in a yard” is “analytic” because it’s trivial—because, basically, it says nothing. But as other philosophers use that word, analytic truths are often decidedly *non*-trivial—for example, “a class is infinitely large exactly if it can be put into a one-one correspondence with a proper subset of itself.” When non-trivial truths are described as “analytic,” what is meant is that they are non-empirically true. (As previously stated, this is what we will mean by it.)

There is one last ambiguity to deal with. As some philosophers use it, the word “analytic” is synonymous with “formally true.” Formal truth is a

property of *sentences*. It's the property of being "syntactically" true. A sentence is "syntactically true" if any sentence having the same syntactic form is true. So "either John is tall or John is not tall" is formally true, and thus analytic (in one sense of the word), because every sentence having the form "either P or not-P" is true. This obviously isn't how we'll be using the term "analytic," given that, as previously stated, we're using it to pick out a property of propositions, not sentences. (Propositions don't have syntax. Syntax is a property of complex *expressions*. Propositions aren't expressions.)

A final note: even though sentence-analyticity diverges from proposition-analyticity, the two ultimately coalesce. What does it mean to say that, although the proposition it encodes is synthetic, a given token of "I am here now" is analytic? It means that, the semantic rules of English being what they are, the proposition that any given token of "I am here now" is true, is analytic. (Given that any token t_1 of "I" refers to the person who utters t_1, that any token t_2 of "now" refers to the time at which t_2 was uttered, and so on, it follows that a token t of "I am here now" couldn't be false.)

4.1 Propositional necessity as microstructural truth (continued)

Analytic propositions are true no matter what. The proposition:

(GG) either grass is green or grass is not green,

cannot fail to be true. It doesn't matter what happens or what natural laws govern what happens. Why is this? What is it about GG that makes it incapable of being false? A natural proposal is to say that GG is *structurally* true. In other words, given what its constituents are, along with how those constituents are put together, it cannot fail to be true. All structural truths are analytic truths. Analytic truth is truth in virtue of structure.

This proposal is almost correct, but not quite. Analytic truth is truth in virtue not just of structure but also of *sub*structure. It isn't merely in virtue of its structure that:

(ACK) any case of knowledge is a case of true belief,

is analytic; the structures of its *constituents* are involved. Some propositions *structurally* just like ACK are false (e.g., any case of knowledge is a case of false belief). At the same time, there are false propositions that are substructurally just like ACK (e.g., "any case of true belief is a case of knowledge"); and it's obviously because its structure differs from theirs that, unlike those propositions, ACK is analytically true. Analyticity is truth in virtue of structure *along* with substructure. Analytic truth is, as we'll put it, high-resolution structural truth.

Before we can see what exactly this means, or why it holds, we must make a few non-controversial, but also non-trivial, points about language. The expressive power of a language is to be understood not only in terms of how many propositions can be expressed, but also in terms of how *varied* those propositions are. A language that can express the propositions *snow is white, it is true that snow is white, it is true that it is true that snow is white,* and so on, can express infinitely many propositions. But its expressive resources are minimal; and some other language that can express only a thousand propositions, not a single one of which entails a single one of the others, will grossly exceed the expressive capabilities of the first. And that's why, when assessing the expressive potency of a language, the relevant question isn't so

much “how many propositions can it express?” as it is “how much redundancy is there within the set of propositions that it can express?”

These points may be obvious, but they have a non-obvious consequence, namely: as a rule, sentential structure mirrors propositional structure; propositions must have structures similar to those of the sentences that represent them. This doesn’t mean that each proposition’s structure is perfectly represented by those of all the sentences that express it. (In fact, this is impossible, given that sentences with different structures can express the same proposition.) It doesn’t even mean that any proposition’s structure is perfectly represented by that of *any* of the sentences that express it. What it means is that, to the extent that the sentences of a given language *fail* to mirror the structures of the propositions they express, that language is expressively frozen. So if *L* and *L** are any two languages, then *ceteris paribus L*’s expressive power exceeds *L**s to the extent that the structures of *L*-sentences do a better job than those of *L**-sentences of mirroring the structures of the propositions they represent.

Let *L** be a language that has only simple symbols. So the *L**-sentence for “snow is colder than water” is some single, syntactically simple symbol, the same being true of every other *L**-sentence. Obviously, there’s no limit to what *L** can express, since there’s no limit in principle to how many primitive symbols can belong to a language. But, since each symbol belonging to *L** is simple, it’s necessary to *add* a new symbol to *L** every time one wants to make a point that isn’t the literal meaning of one of its simple expressions. Nothing could be said in *L** that hadn’t been said before: it would be impossible to express *discoveries.* Symbols would have to be added *ad hoc* to say anything new. But given a language *L* that has many ways of combining its primitive symbols into complex ones, it’s expressive capabilities will dwarf those of *L**, so long as the number of primitive expressions belonging to *L* rises above a certain, extremely small minimum number. *L** is totally inflexible. *L* is highly flexible, and can do a lot with a little. Other things being equal, the extent to which one language’s expressive resources exceed those of another is directly proportional to the extent by which sentences of the one language exceed their counterparts in the other in respect of syntactic complexity.

There is only one viable explanation of this last fact: Syntactic structure mirrors propositional structure. The syntactic devices that endow a language with expressive flexibility correspond to the relations by which propositions

are constructed out of their constituents and, in particular, by which propositions are constructed out of other propositions. Thus, we can confidently say that the (syntactic) structures of sentences tend to mirror those of the propositions they express.

4.2 Necessity as high-resolution structural truth

We can now say what it means to say that analytic truth is high-resolution structural truth. Given the reasoning just presented, the sentence "either grass is green or grass is not green" must have a structure similar to GG. But, by that same reasoning, GG's structure probably differs from the structure of:

(GA) Grass is green and grass is not green.

But GA is not analytically true. (It's analytically false.) Given that GA is false and has the same structure as GG, it cannot possibly be solely in virtue of its structure that GG is true. But we don't just want to give up on the idea that analytic truth is structural truth. An analytically true proposition is *ipso facto* true no matter what contingencies obtain. This entails, or strongly suggests, that it is always in virtue of something *inherent* in it that a proposition is analytically true, and it's hard to see how the relevant property could fail to be structural in nature. It's obvious that, in the case of GA, what's causing the trouble is that the concept of conjunction (the concept of *and*) is occupying the place that, in GG, is occupied by the concept of disjunction (the concept of *or*). Taken together, these points suggest that it is in virtue of its structure *along with* the structure of one of its constituents—viz., the concept of conjunction—that GG is analytically true.

So it's not quite right to say that analytic truth is structural truth. At the same time, it isn't quite right to identify analytic truth with *sub*structural truth. The proposition:

(NG) it's not the case that grass is green or grass is green,

is equivalent with the proposition that grass is not green. Thus, NG is contingently false. But it has the same constituents as the analytically true GG, and is thus substructurally identical with it, even though it's structurally very different.

This suggests that for a proposition to be analytically true is for it to be true in virtue of its structural *and* its substructural properties. In other words, given a proposition's structure along with the structures of its constituents, it's settled whether or not that proposition is analytically true. Thus, a *high-resolution* representation of a proposition's structure settles whether or not

it’s analytic. (A “low-resolution” representation would be one that failed to disclose substructure.) Thus, *analytic truth is high-resolution structural truth* —and so, therefore, is necessary truth, given that it’s identical with analytic truth.

5.0 Quine's denial of informal analytic truth: preliminary concepts

In this section, all references to "analytic" truth are to sentential analytic truths; we will be discussing analytically true sentences, not analytically true propositions.

W.V.O. Quine (1951) says that, with a few trivial exceptions, there is no such thing as sentential analytic truth. To understand exactly what it is that he holds, and why he holds it, we must first make some background points.

Consider the following sentence:

> (ST) "if Smith is tall, then it is not the case that Smith is not tall,"

Obviously you can't replace "Smith" with "therefore" or "red." The result would be syntactically ill-formed nonsense. Thus, such replacements would not be "grammatically admissible." But you *can* replace "Smith" with "Jones" and "that person." Such replacements *would* be grammatically admissible, that being why:

> (JT) "if Jones is tall, then it is not the case that Jones is not tall,"

is a perfectly grammatical sentence.

JT is what results when the occurrences of "Smith" in it are replaced *uniformly.* In other words, JT is what results when not just one, but *all* of those occurrences are replaced with the same thing. The following sentence, by contrast, is what results when only *one* of the occurrences of "Smith" in ST is replaced with "Jones" ("NU" is short for "not uniform"):

> (NU) "if Jones is tall, then it is not the case that Jenkins is not tall."

Thus, a syntactically well-formed sentence *S* cannot be false if it is the result of uniformly replacing the occurrences in *ST* of "Smith" (or, for reasons analogous to those just given, of "tall") with some other expression.

In general, a sentence *S* is *formally* true if there is some expression *E* occurring in *S* such that no syntactically well-formed sentence *S** is false if

*S** is what results when the occurrences in *S* of *E* are uniformly replaced with occurrences of some other expression.

It's generally held—and the present author himself holds this—that there are *informal* analytic truths. Consider, for example, the following sentence ("JTB" is short for "justified true belief"):

(JTB) "One can't know something without believing it."

JTB is analytically true. Given *only* a knowledge of what it means, one knows (or is in a position to know) that it is true. It isn't like "there are nine planets," which, if true, can be established only on the basis of experimental data. No—JTB is *inherently* true. The denial of it ("one *can* know something without believing it") is prevented, by its very own meaning, from being true; and it (JTB) is *required* by its very own meaning to be true.

But JTB isn't *formally* true. If you replace "know" with "doubt," the result is a false sentence. In fact, given *any* expression in JTB, replacing it (in a uniform and grammatically admissible manner) with some other expression results in falsehood. For example, the sentence "one can't deny something without affirming it" is a false sentence that results from replacements of the sort just described.

It seems pretty clear, then, that there are informal analytic truths. And this is what most analytic philosophers hold, including the present author. But Quine denies it. Quine thinks that, apart from formal analytic truths, there are *no* analytic truths. This thesis of his is famous, and so is his argument for it.

5.1 Quine's denial of informal analytic truth

Here is Quine's argument. Consider the sentence:

(B) Bachelors are unmarried men.

If any sentence is analytic, (B) is one of them. According to Quine, supposing for argument's sake that (B) is analytic, the only possible reason is that "bachelor" and "unmarried man" are synonyms. Quine holds that *all* analytically true statements, supposing such things to exist, merely register synonymy-relations and, therefore, that all of them are as trivial as (B).

Quine then asks the following question: What is it for two expressions to be synonyms? Quine gives the following (correct) response: It is for them to be such that intersubstitutions of them preserve meaning. "Foe" and "enemy" are synonymous because you cannot change sentence meaning by replacing "foe" with "enemy." Put another way, "foe" and "enemy" mean the same thing because, if a sentence *S* contains an occurrence of "foe," and that occurrence is replaced with an occurrence of "enemy," the resulting sentence has the same meaning as *S*. Consider the sentence: "Smith is a foe." Now replace the word "foe" with "enemy." What results is "Smith is an enemy." The sentence "Smith is an enemy" has precisely the same meaning as "Smith is a foe."

This line of thought is easily extended. Let *S* and *S** be *any* two sentences that are exactly alike *except* for the fact that, in the place where *S* contains the word "foe," *S** contains the word "enemy." What we just said about "Smith is an enemy" and "Smith is a foe" is true of *S* and *S**. That is why "foe" and "enemy" are synonyms. Inter-substituting synonyms doesn't change sentence-meaning—and for two expressions to be synonymous just *is* for them to be such that inter-substituting them doesn't ever change sentence-meaning.

But, Quine asks, what is it for two sentences to have the same "meaning"? What is it for "Smith is an enemy" and "Smith is a foe" to mean the same thing? According to Quine, for two statements to have the same meaning is for them to be *analytically equivalent.* In other words, it is for them to entail, and be entailed by, exactly the same statements. "Smith is an enemy" entails "Smith is somebody who would like to do me ill and will do so if given the opportunity," and so does "Smith is a foe." The statement "Smith is

somebody whose sole objective in life is to make my life as unpleasant as possible" entails "Smith is an enemy," and it *also* entails "Smith is a foe." Given *any* statement *S* that entails "Smith is an enemy," *S also* entails "Smith is a foe"; and given *any* statement *S** such that "Smith is an enemy" entails *S**, "Smith is a foe" *also* entails *S**. Thus, "Smith is an enemy" and "Smith is a foe" entail, and are entailed by, exactly the same statements—and that, according to Quine, is what it is for them to have the same meaning.

But notice that we have gone in a circle. We defined analyticity in terms of synonymy. A recap will make this clear:

> 1. (B) is analytic because "bachelor" and "unmarried man" are synonyms, and there is no analytically statements that isn't to be understood along similar lines: all such statements are sterile trivialities that hold in virtue of linguistic conventions.

We then defined synonymy in terms of sameness of meaning:

> 2. For two expressions to be synonymous is for them to be such that inter-substitutions of them preserve sentence meaning.

But we then defined sameness of meaning in terms of analyticity:

> 3. For two statements to have the same meaning is for them to be *analytically equivalent.* In other words, it is for them to entail, and be entailed by, exactly the same statements.

1 defines analyticity in terms of 2 and, thus, in terms of synonymy. 2 defines synonymy in terms of 3, and thus in terms of analyticity. Thus, analyticity has been defined in terms of itself. We have gone in a vicious circle. Thus, Quine concludes, the concept of analyticity is viciously circular and therefore incoherent. Quine concludes that there is no such thing as analytic truth: all truths are synthetic.

5.2 Evaluating Quine's argument

First of all, Quine is wrong to hold that analyticity is to be understood in terms of synonymy. Consider the sentence:

> (T) The unique number *n* such that *n* = 2 if the interior angles of a Euclidean triangle add up to 180° and *n* = 3 otherwise is identical with the unique number *m* such that 9^n = 90 minus 9.

(T) is analytic. No empirical knowledge, over and above such as is needed to know what meaning it bears, is required to know whether it's true. So it's true in virtue of what it means and is thus analytic. (To be sure, unlike some analytic sentences, it isn't *formally* true. But since the very thing that Quine is trying to show is that there is no informal analytic truth, he cannot, without assuming the truth of the very thing that he's trying to show, claim on those grounds that T is not analytic.) So if Quine is right, T holds in virtue of some synonymy relation. What, in this case, would the relevant synonyms be? Presumably they would be:

> (*) "the unique number *n* such that *n* = 2 if the interior angles of a Euclidean triangle add up to 180° and *n* = 3 otherwise,"

and:

> (#) "the unique number *m* such that 9^m = 90 minus 9."

But even though they *refer* to the same thing (namely, the number two), * and # obviously don't have the same *meaning;* and Quine is therefore wrong to say that, if there are analytic truths, they always hold in virtue of synonymy relations.

5.3 Analytic truth non-circularly definable as non-empirical truth

Quine's larger point seems to be that there is no non-circular definition of the term "analytic truth." But that is false: a sentence is analytic if (a) it is true and (b) it does not hold in virtue of *empirical* facts. Put simply: analytic truth is non-empirical truth. "Circles are figures of uniform curvature" is true and, moreover, it holds, not in virtue of empirical facts (i.e., not in virtue of how objects are *in fact* shaped), but in virtue of facts about the structures of the properties *circle*, *curvature*, and so on. Of course, being a nominalist, Quine would reject the idea that there are non-empirical truths. (After all, a nominalist believes that everything is spatiotemporal: there are no platonic entities and thus no properties.) But in attempting to establish that all truths are synthetic, Quine is, in effect, trying to *establish* that there are no empirical truths. For an empirical truth is, by definition, one that can be learned only through observation and therefore cannot be learned through any non-sensory modality such as conceptual analysis. So Quine would be guilty of begging the question were he to say that, since all truth is empirical, it is not an option to define "analytic" as "non-empirically true."

5.4 Quine's position self-defeating

Also, the view that there are no analytic statements is self-defeating.[226] The statement:

(^) There are no empirical statements,

is equivalent to the statement:

(^^) for all values of *S*, the statement *nothing follows from S* follows from the statement *S is a statement.*

In other words, if you say that there are no analytic statements, you are saying that nothing follows from any statement. But if you say *that,* then you are saying that the statement *nothing follows from S* follows from the statement *S is a statement.* Thus, the view that there is no analytic truth entails that *there is* analytic truth, and is therefore self-refuting.

5.5 Why Quine's position entails that nothing can confirm anything

There is another reason to hold that, contrary to what Quine says, there are in fact analytic truths. To say that there is no analytic truth is to say that all truths are empirical truths. For the sake of argument, let us suppose that all truths are indeed empirical. In that case, given any two statements S_1 and S_2, it would be an *empirical* question—one to be decided on observational grounds—whether S_1 supported S_2; and, supposing that S_1 did support S_2, it would be an empirical question to what degree it did. This means that, for any values m_1 and m_2, and for any observations $A_1 \ldots A_n$, it would be an empirical question whether or not:

> (S_3) Observations $A_1 \ldots A_n$ confirm to degree m_1 that S_1 confirms S_2 to degree m_2,

is correct. Further, it means that, for any value m_3, and for any observations $B_1 \ldots B_o$, it would be an empirical question whether or not:

> (S_4) observations $B_1 \ldots B_o$, confirm to degree m_3 that observations $A_1 \ldots A_n$ confirm to degree m_1 that S_1 confirms S_2 to degree m_2,

were correct.

And so on *ad infinitum.*

Thus, the supposition that all truth is empirical is viciously regressive. Thus, if there is no analytic truth—that is, if all truth were empirical---then nothing could confirm anything and, as Laurence Bonjour (1998) puts it, all inquiry would "grind to a halt."

6.0 Why there are necessarily true propositions

First a terminological point: If a proposition is true, but it isn't necessarily true, it is "contingently true," (or just "contingent"); and if a proposition is false, but it isn't necessarily false, it is "contingently false."

It seems clear that *some* truths are contingent. Gore *might* have become President in 2001. He didn't. Bush became President instead. But Bush didn't *have* to become President. So it's a contingent truth that Bush became President. In general, things don't *have* to be the way they are; they could be different. In any case, this is what we think; and it's what we'll assume in this context.

Given this assumption, we must ask the following question: Are any propositions necessarily true? Yes. Why is this? Because the assumption that no propositions are necessarily true is self-undermining. For argument's sake, suppose that:

(NN) No propositions are necessarily true.

Is NN itself necessarily true or not? It has to be one or the other. If NN is a necessary truth, it's a counterexample to itself and is therefore false. So let's suppose for argument's sake that NN is contingently true. In that case, it *could* be false. In other words, there is some possible world *W* where some proposition *P* is necessarily true. A proposition is necessarily true iff it's true in all possible worlds. So given that *P* is necessarily true in *W*, it follows that *P* is true in *every* world. Thus, *P* is true in our world, and *P* is true in every world. Thus, *P* is *necessarily* true in our world, and NN is therefore false. To sum up: NN is self-defeating if it's necessarily true, and NN is self-defeating if it's true but not necessarily so. Therefore, NN is false.

A similar argument: if NN is true, then the statement (i) "*P* isn't necessarily true" necessarily follows the statement (ii) "*P* is a proposition," and it is therefore a necessary *truth* that (i) isn't false if (ii) is true. So NN cannot be true unless it's false, and it's therefore false.

6.1 Knowledge of necessary truth is a prerequisite for knowledge of contingent truth

Nonetheless, some philosophers have denied that there are necessary propositions. We just considered Quine's argument for this. Given how contrived that argument is, and how lacking in logical integrity and devoid of intuitive force, it's hard to believe that Quine's rejection of necessary truth was based entirely on it. Quine, one would suspect, must have had some pre-existing agenda and he constructed that argument in the hopes that it would legitimate that agenda.

Such suspicions would be correct. Throughout his long career, Quine was unrelenting in his advocacy of a particularly extreme form of empiricism.

To understand Quine's empiricism, we must note a fact about empiricism in general. Empiricism says that you know only what your senses tell you. (More precisely, empiricism tells you that, if you know it, then either (a) you either sense-perceived it *or* (b) you sense-perceived direct evidence of it, where *x* is direct evidence of *y* if sense-perception by itself is sufficient to apprise one of the inference rule whereby, given *x*, *y* can be inferred.) Unless you assume the truth of principles that cannot be learned empirically, you can't show that your perceptions correspond to any transperceptual reality. For this reason, empiricism tends to collapse into idealism, this being why Berkeley, who has a hardcore empiricist, was also an idealist.

Quine's empiricism is antithetical in at least one respect to Berkeley's. Berkeley was led by his empiricism to deny the non-mental. Quine was led by his empiricism to deny the mental. According to Quine, there is no observational evidence of mental states. One sees bodies. One *doesn't* see or otherwise sense-perceive mental states. Quine's position is that mental states either don't exist or are identical with the bodily behaviors that we ordinarily, and wrongly in Quine's view, believe to be *expressions* of mental states. Thus, Quine was a behaviorist. (In Chapter 22, it is shown why behaviorism is false.)

Quine also held that there is no *a priori* knowledge. There is, he believed, no knowledge that is hardwired into us and that regulates the intake and subsequent processing of sensory information. Quine also believed that our linguistic behavior can be explained on the assumption that, at birth, we are blank slates and that it is entirely through a process of Pavlovian conditioning

that we come to use words in the way we do.

This doctrine is *also* known as "behaviorism." Behaviorism in *this* sense is a psychological theory that tries to answer the question "how does a human being learn to speak a language?" Behaviorism in the other sense is a philosophical theory that tries to answer the question "what are mental states?" To avoid confusion, I'll refer to the first doctrine as "linguistic behaviorism."

Linguistic behaviorism is incoherent. To say "that's a fire" and *mean* it, I must have beliefs as to what that expression means and as to what circumstances would warrant my uttering it. If I reflexively bark it out when I see a fire, I'm no more affirming that there's a fire than I am if I start sweating in response to the heat given off by a nearby fire. These points are substantiated in Chapters 1, 3, and 8.

Also, in 1957, Noam Chomsky published his groundbreaking work *Syntactic Structures*, which, along with the many other important works that Chomsky was soon to publish, made it clear that Quine's linguistic behaviorism simply didn't model the relevant facts. Chomsky proposed an alternative theory, according to which inborn cognitive structures are largely, though obviously not entirely, determinative of how we learn and use language. Unlike Quine's theory, Chomsky's is splendidly borne out by the facts.

Quine and others accused Chomsky of putting forth *a priori* theories that lacked an adequate empirical grounding. This is both false and confused. It's true that *according* to Chomsky's theory we have *a priori* knowledge. But it doesn't follow that Chomsky's theory is itself *a priori*. Chomsky's theory is based on empirical data; and, unlike Quine's theory, it fits that data. And even though, according to Quine's theory, there is no *a priori* knowledge, it doesn't follow that Quine's theory is itself empirically well-founded. It isn't. It's rooted in the *a priori*, and easily disconfirmed, presumption that all psychological phenomena are to be explained in the same way as the salivating of Pavlov's dog.

One of the great ironies of empiricism is that it's one of the most unempirical theories of all time. People learn languages. Dogs don't. A dog wouldn't even learn the rudiments of a language were it to experience the sensory simulations that enable a human being to master one. This shows that, unlike dogs, human beings are endowed with innate cognitive structures that enable them to derive an integrated and complex body of knowledge

from such stimulations. The same point *mutatis mutandis* holds of any cognitive skill that human beings can acquire that animals cannot. Empiricism is therefore inconsistent with empirical evidence of the most blatant and widespread kind.

Despite the vast and ever-growing body of evidence in support of Chomsky's anti-Quinean views, Quine stuck to his guns, and in the 45 years that passed between the publication of *Syntactic Structures* and his death, Quine didn't soften his position. Not only did he not concede that Chomsky was at all right; he didn't even concede that mental entities even exist, except to the (non-existent) extent to which they can be identified with overt behaviors.[227]

In any case, it's clear that Quine's rejection of analytic truth wasn't the effect, but was instead the cause of his acceptance of the bizarre and unwieldy argument of his that we considered earlier. Quine obviously produced that argument *post hoc*, hoping to give his idiosyncratic form of empiricism a solid foundation. Quine's real reason for denying the existence of analytic truth lay in his staunch empiricism. All philosophers, and indeed non-philosophers, who deny the existence of analytic truth do so because they are staunch empiricists.

Thus, when authors deny that there are necessary truths, the underlying sentiment, if not the immediately operative rationale, is this:

> (NT) If they exist, necessary truths hold no matter what. They hold no matter how the world is and, therefore, no matter what observations one has. So, if they exist, they can be known *independently* of observation. But it's unscientific to think that anything could be known in any way other than through empirical observations. Therefore, there are no necessary truths.

6.2 Why NT is a spurious argument

All knowledge of the spatiotemporal world is rooted in observation. But there's more to doing science than making observations. The observations are only a means to an end, the end being the acquisition of theoretical knowledge. We'll now see that non-observational knowledge is a precondition for theoretical knowledge and also that it is always necessary truths that are known non-observationally.

Our senses give us raw data. Raw data is the foundation of all empirical knowledge. But data cannot interpret itself. And even if it could, it wouldn't relieve you of the burden of having to interpret the data for yourself. For argument's sake, suppose that the data were self-interpreting, and that its self-interpretations were disclosed to you in periodic reports. Any such report would *itself* be given to you in terms of raw data that you'd have to analyze, a consequence being that you couldn't even understand the report unless you already knew how to interpret the data.

Second, once you did understand the report, you'd have no good reason to trust it unless you could find some independent corroboration for it. Finding such corroboration would involve coming up with your own interpretation of the data and seeing how well it matched the interpretation in the report. So even if *per impossible* perceptual data could self-interpret, one could learn nothing about the external world from those interpretations unless one could interpret the data for oneself.

It is, therefore, totally incoherent to suppose that one could know on the basis of observation alone how to interpret the data of observation; and since it's totally incoherent to suppose that one could have theoretical or, therefore, scientific knowledge without interpreting the data, it is totally incoherent to suppose that any science should be *strictly* observation-based. There is necessarily a non-observational component to the acquisition of scientific knowledge.

Interpreting data involves being able to discern abstract resemblances between different concrete situations. Doing this, in its turn, involves knowing necessary truths. Seeing that two objects have the same shape involves seeing that they fulfill conditions that, if *not* fulfilled by a given object, necessarily preclude it from having that shape. Seeing that otherwise dissimilar objects *x* and *y* are both square-shaped involves knowing that, were

an object to have more sides than *x* or *y*, it would *for that very reason* fail to have the same shape as them, and it also involves knowing that, if an object had a different color from *x* or *y*, it would not, at least not *for that very reason*, fail to have the same shape as *x* and *y*. One can't know that *x* and *y* have the same shape unless one knows that an object's not having four sides *necessarily* disqualifies it from being shape-identical with those objects. In general, one cannot discern *any* abstract similarities between *any* two concrete situations unless one knows, for at least some property, what conditions *must* be fulfilled if a given object is to have that property. So knowledge of necessary truths is a precondition for being able to recognize resemblances between things. Since knowledge of such resemblances is the first, and the most rudimentary phase, of scientific endeavor, it is radically incoherent to suppose that there should be any scientific knowledge without knowledge of at least some necessary truths.

W.V.O. Quine tried to provide a rigorous proof of NT. We considered this proof in Sections 5.0–5.5 of the present chapter and found it to be wanting. But notice that the very idea of trying to *prove* that contention is a self-defeating one. For there are no proofs if all truth is contingent, since proving something is showing that it *must* be the case. (More accurately, to prove conclusion *C* on the basis of premises *P* is to show it to be necessarily true that if *P*, then *C*.) Thus, to prove NN to be true is to prove it to be false. So a certain incoherence is embedded in the very idea of trying to show that there are no necessary statements.

7.0 Analytic truth vs. formal truth

Formal truth isn't analytic truth. We'll discuss this in this section, and we'll also see what exactly it means to describe a truth as "formal." A statement is analytically true if its negation is incoherent. "Squares have four sides" is analytic. Why? Because "squares don't have four sides" is incoherent, and so is "anything that is literate is inanimate." The reason being that it makes no sense to suppose that squares might fail to have four sides or that literate beings might fail to be animate.

"Right now, in 2009, Barack Obama is the U.S. President" is a non-analytic truth. This is because "Barack Obama is not President in 2009," though false, makes sense. If that sentence didn't make sense, then in 2007 it wouldn't have been reasonable to wonder whether Obama would be President in 2009. But it was.

Non-analytic truths are ***empirical*** truths. An empirical truth is one that can be known only through sensory observation. It is only on the basis of sense-perception that one can know that Barack Obama is currently the U.S. President. That is why "Barack Obama is currently the U.S. President" expresses an empirical truth. But it is entirely through insight into its meaning, and into its *entailment-relations,* that one knows that nothing could fail be to be identical with itself; that being why that truth is analytic.

7.1 Empirical vs. pseudo-empirical truth

One must distinguish genuine empirical knowledge, the acquisition of which *coincides* (at least during its initial stages) with the sense perception, from analytic knowledge, the acquisition of which is merely *triggered* by sense perception. Consider the principle that:

> (DN)[228] if a statement is false, then the negation of its negation is false.

For example, if Smith is in Richmond, then it is false that Smith is not in Richmond. DN is one of the cornerstones of logic.

Learning DN may involve hearing or reading, and therefore seeing, some affirmation of it. But you don't learn DN empirically—that is, it isn't through sensory observation that you learn it. But, if you know DN, it is because you see that, given its meaning, it would make no sense to deny it. Sense-perceptions may well have been needed to *initiate* or *trigger* the reflections on the basis of which you learned DN. But such sense-perceptions were themselves no part of that discovery process. They were part of the *pre*-discovery process, much as turning the ignition in a car is part of the *pre*-journey, as opposed to the journey proper.[229]

Given a true statement *S*, there is a simple test for determining whether it is empirical or analytic. If, in order to *justify* acceptance of *S*, you must cite empirical data, it's empirical. To justify your acceptance of the statement that Smith has gained 20 pounds in the last year, you must cite empirical data (e.g., sense-perceptions that you have had of Smith or of affirmations on the part of reliable sources as to Smith's weight situation). But your acceptance of DN can be justified only by insight into its meaning; it cannot possibly be justified on observational grounds. Of course, one might justify one's acceptance of DN by citing the fact that one has heard some authority affirm it; and, of course, one's knowledge of that fact is observation-based. But those observations justify DN *only* to the extent that the authority figures in question understand why, for reasons that lie entirely in DN's content and have nothing to do with anything one perceives, that statement must be true. Testimony-based knowledge of analytic truths, if it exists, is parasitic on non-observational knowledge of them.

And it's doubtful whether such knowledge really exists. If I don't

understand why DN is true, and I accept it only because somebody I trust does so, then I don't really know it. What *I* accept is not, it seems to me, quite what the authority figure in question accepts. That person really accepts DN. But what I accept isn't DN, and is instead some statement along the lines of: "since so and so is authoritative, and since so and so affirmed DN, that statement is probably right"—which, unlike DN itself, is empirical.

Sense-perception adds to one's knowledge in two different ways. In some cases, it does so by *transmitting* information about the external world. If I see Jim playing soccer, my visual perception adds to my knowledge by transmitting to me the correct message that Jim is playing soccer. In other cases, sense-perception merely *activates* some purely ratiocinative process. The ratiocination-activating role of sense-perceptions must be distinguished from its information-transmitting role.

7.2 Analytic knowledge of principles vs. empirical knowledge of the fact that certain sentences express those principles

It's an empirical fact that the *sentence* "squares have four sides" means what it does. Those sounds could mean anything. They could mean that squares don't have four sides or that everybody owns a Rolls Royce. A brilliant extraterrestrial could reasonably wonder what that sentence means. But such a being could *not* reasonably wonder whether squares have four sides. Thus, it's an *empirical* truth that the *sentence* "squares have four sides" is to the effect that squares have four sides. But it's an analytic, non-empirical fact that squares have four sides.

7.3 Analyticity vs. triviality

The concept of analytic truth is typically *illustrated* in connection with trivial statements (e.g., “any given thing is identical with itself”). But there are ***non-trivial*** analytic truths. An analytically true statement is trivial if one cannot know what it means without knowing that it's true. (So "Smith is identical with Smith" is both analytically true and trivial.) An analytically true statement is non-trivial if one *can* know what it means without knowing whether it's true. Here are some examples of non-trivial analytic truths:

> * Given only that an institution is flourishing, it cannot be inferred that the individuals composing it are flourishing; and given that the individuals composing an institution are flourishing, it cannot be inferred that the institution itself is flourishing.
>
> *No finite set can be put into a one-one correspondence with a proper subset of itself.
>
> *Any infinite set can be put into a one-one correspondence with a proper subset of itself.
>
> *Anything sentient is, at any given point, unaware of at least some of its mental states.
>
> *There are cases of justified true belief that are not knowledge.
>
> *There are continuous functions that cannot be differentiated at any point.
>
> *Two objects that cannot enter into causal relations with each other cannot have spatial or temporal positions relative to each other.

Also, it’s an analytic truth, and also a non-trivial one, that there are non-

trivial analytic truths. We saw this in Chapter 1.

7.4 Formal truth

The concept of *formal truth* has an important place in philosophical inquiry. It is sometimes propositions that are described as “formally true.” But it is almost always linguistic expressions that are so described. To the extent that formal truth is a property of expressions, it must be a property of sentence-*tokens*, not of sentence-types, since the latter, as we saw in Chapter 3, are never true or false.

To avoid generating too much verbal clutter, I'll sometimes use the word "sentence" to mean *sentence-token.* But it mustn't be forgotten that no linguistic expressions other than sentence-tokens are capable of being true or false.

Here are some examples of sentence-types (note the just-made disclaimer) whose tokens would universally be agreed to be “formally true”:

> (1) If Bill is tall, then it is not the case that Bill is not tall.
>
> (2) If everything has a given property, then any given thing has that property.
>
> (3) If all F’s are G’s, and no G’s are H’s, then no F’s are H’s.
>
> (4) Either Helsinki is in Finland or Helsinki is not in Finland.

What is “formal truth”? What does it mean to describe (1), for example, as “formally true”? Here’s the canonical answer (“FT” stands for “formal truth,” of course):

> (FT) Formal truth is *syntactic* truth. A sentence is “syntactically true” if, given only its syntactic structure, it cannot fail to be true. Thus, a sentence is formally true if no sentence having the same syntactic structure is false.

FT cannot be accepted. A linguist would say that (1) “grass is green or grass is not green” has the same syntax as (2) “grass is green and grass is not green.” As linguists use the word “syntax,” two expressions have the same syntax if the way in which the one is built out of its parts is identical with the

way in which the other is built out of *its* parts. For example, “Bill hates Ted” and “Laura loves Sally,” though having very different meanings, have the same *syntax*, the reason being that they’re structurally identical. The one is built out of its simple constituents in the very way that the other is built out of *its* constituents. Thus, it isn’t because of any structural differences between them that they have different meanings.

The relevant difference is *sub*structural. It lies in the fact that the *meaning* of “and” is different from that of “or.” So the difference between (1) and (2) is like the difference between “John is tall” and “Mary is short”: it lies entirely in the fact that the primitive, semantically unstructured constituents of the one have different meanings from their counterparts in the other. Thus, relative to what *linguists* mean by “syntax” (and “syntactic,” etc.), it’s a non-starter to say that “formal truth” is syntactic truth. Logicians and philosophers obviously have in mind some very different understanding of the word “syntax” when they identify “formal” truth with “syntactic” truth. They *don’t* have in mind sentential (or phrasal) structure. What *do* they have in mind? What do philosophers mean when they say that a formally true sentence is one that holds entirely in virtue of its “syntax”? What do they mean when they describe some statement as a “formal truth”?

7.4.1 Quine on formal truth

W.V.O. Quine (1908–2000) gave a characteristically intelligent and straightforward answer to this important question: a sentence S is formally true if, given any expression in S that isn't a *logical constant*, it isn't possible to produce a false sentence by replacing that expression with another expression of the appropriate grammatical category. A "logical constant" is an expression on the following list: "if . . . then . . . ," "either . . . or . . . ," "not," "for all," "for some," "="—and any obvious derivatives of any of these expressions (e.g., "is identical with," "given any," "it is not the case that," etc.)[230]

Consider the sentence "grass is green or grass is not green." Replacing "grass" with "fire" or "Bob" results in a truth, as does replacing "green" with "cool" or "voluminous." (It's assumed that the replacements are uniformly made: if a single occurrence of "grass" is replaced with "fire," so is every other such occurrence.) Given any expression in (1) *other* than the logical constants in it (viz., "either . . . or . . ." and "not"), replacing that expression with another belonging to the same grammatical category results in truth.

There are two problems with Quine's analysis. First, it doesn't say what is so special about those items; it doesn't say what these so-called "logical constants" have in common with one another that other expressions *not* on the list *don't* have in common with them. And until that information is provided, there's nothing to prevent one from extending that list in arbitrary ways. Unless that list was generated by some rule, anything can be added to it without diminishing its integrity, since it has no integrity to begin with. But if it was generated by some rule, then Quine's answer fails to identify it and thus fails to say what formal truth is.

An analogy may help. When asked to define the expression "number," I simply *list* the various different kinds of numbers ("real," "whole," "rational," "imaginary," etc.). Have I answered the question? No. Even though I've listed the various different kinds of numbers, I haven't said what numbers are, and I thus haven't provided any *reason* why fractions are on that list but birds are not. Accordingly, my answer doesn't give anyone a reason not to extend that list as he or she pleases—not to decide that birds and rocks are numbers. Of course, nobody *would* extend the list in such an absurd way. But that's only because our intuitions strongly suggest that there is *some* characteristic had in common by reals, rationals, etc., that birds and

rocks don’t have.

Second, there are truths not covered by Quine’s analysis that don’t appear to differ in any relevant respect from those covered by that analysis. Many modal statements are in this category—for example, “if it’s necessarily the case that Smith is a mammal if Smith is a dog, then if it’s necessarily the case that Smith is a mammal if it’s necessarily the case that Smith is a dog.” There doesn’t seem to be any relevant difference between *that* statement and one that conforms to Quine’s analysis—for example, “if Smith is a mammal if Smith is a dog, and Smith is an animal if Smith is mammal, then Smith is an animal if Smith is a dog.” So there doesn’t seem to be any good reason why “necessarily,” unlike “if . . . then . . . ,” doesn’t make it onto Quine’s list.

Of course, one could deal with this by adding “necessarily” to that list. But then we’d have to ask: “what makes *that* extension of the list legitimate?– bringing us back to the criticism made a moment ago.

7.4.2 Why formal truth ≠ topic-neutral truth

Some have argued that formal truth is maximally generic, or "topic-neutral" truth. The idea is that "$x = x$" is logically true because it's completely general: it doesn't belong to any one discipline and is presupposed by all of them.[231]

But this doesn't work. There are maximally generic truths that aren't formal—for example, "nothing can wholly pre-exist itself," "nothing spatiotemporal can lack causal properties," "nothing non-spatiotemporal can have causal properties," "inductive inference isn't an appropriate method of reasoning for non-empirical disciplines." Given any one of these sentences, it's easy to show that there are false sentences having the same syntactic form. For example, the first of those sentences has the same form as "nothing spatiotemporal can have causal properties," which is obviously false.

One could argue that these truths *aren't* sufficiently generic. But then it would be very unclear what exactly "sufficiently generic" meant in this context, and the suspicion would arise that, so far as it had any coherent meaning, "sufficiently generic" meant "not a counterexample to this particular theory of formal truth."

Also, whether a sentence expresses a generic truth has nothing to do with its form. The truth expressed by "if everything has a given property, then any given thing has that property" is surely a maximally generic one. But that truth could obviously be expressed by a primitive, syntactically unstructured symbol. In fact, there is *no* form that sentences must have if they are to express that truth or any other. So, if "formal truth" is taken to be a characteristic of *expressions* of some kind, then it's simply false to identify "formal truth" with "maximally generic" truth. And if "formal truth" isn't meant in this way, then it's simply irrelevant to our current inquiry.

7.4.3 An alternative analysis of formal truth[232]

A sentence is formally true if it is an instance of an open-sentence that is true for all values of its free variables. (An "open-sentence" is an expression that contains a free variable, and is therefore neither true nor false, but is otherwise just like a sentence. So "x is tall" is an open-sentence.[233]) For example:

(5) If Smith is at home, then it's not the case that Smith is not at home;

is a formal truth because it's an instance of the open-sentence *if x is phi, then it's not the case that x is not phi*, which is true for all values of *x*. (We'll see that (5) is an instance of many such open-sentences. But its being an instance of just one is enough for its being formally true.) Given any formal truth, the same holds.

If an open-sentence is true for all values of its variables, we'll say that it's "universally true." So "$x = x$" is universally true (that being why "Smith = Smith" is "formally" true). It must be kept in mind that this is just an abbreviation. "$x = x$" is neither true nor false. No open sentence is true or false.

In light of these points, consider the following expressions:

(1) Bill Gates is wealthy.
(2) *x* is wealthy.
(3) Bill Gates is phi.
(4) *x* is phi.

(1) is a true sentence. (2)–(4) are neither true nor false. Given (1), (2) is what results when "Bill Gates" is replaced with a variable, (3) is what results when "wealthy" is replaced with a variable, and (4) is what results when both are replaced with variables. (See Chapter 7.) Of course, actual sentences are formed out of these open-sentences by replacing the variables in (2)-(4) with constants. In each case, this procedure may yield either a true sentence or a false one. For example, replacing the variable in (2) with "Bill Gates" yields a truth and replacing it with "Mike Smith," the name of an impecunious friend of mine, yields a falsehood.

Bearing these points in mind, consider the following expressions:

(5) If Smith is at home, then it's not the case that Smith is not at home.
(6) If *x* is at home, then it's not the case that *x* is not at home.
(7) If Smith is phi, then it's not the case that Smith is not phi.

(8) If *x* is phi, then it's not the case that *x* is not phi.
(9) If *P*, then it's not the case that not-*P*.

Each of (6)–(9) is formed by replacing one or more expressions in (5) with a variable. And the following sentences result when, in each case, the variables are bound by universal quantifiers.

(6_{UG}) For any individual *x*, if *x* is at home, then it's not the case that *x* is not at home.
(7_{UG}) For any property phi, if Smith is phi, then it's not the case that Smith is not phi.
(8_{UG}) For any individual *x* and any property phi, if *x* is phi, then it's not the case that *x* is not phi.
(9_{UG}) For any proposition *P*, if *P*, then it's not the case that not-*P*.

This suggests that a sentence is formally true if it's an instance of a universally true open-sentence.

To take another example:

(10) If it's necessarily the case that Smith is a mammal if Smith is a dog, then it's necessarily the case that Smith is a mammal if it's necessarily the case that Smith is a dog;

is formally true, because it's an instance of the following universally correct open-sentences:

(11) If it's necessarily the case that *x* is a mammal if *x* is a dog, then it's necessarily the case that *x* is a mammal if it's necessarily the case that *x* is a dog.

(12) If it's necessarily the case that Smith is a mammal if *P*, then it's necessarily the case that Smith is a mammal if it's necessarily the case that *P*.

(13) If it’s necessarily the case that *Q* if *P*, then it’s necessarily the case that *Q* if it’s necessarily the case that *P*.

7.4.4 What is a logical constant?

Quine's analysis, we noted, is vitiated by his failing to say what conditions an expression must satisfy to be a "logical constant." We are in position to provide this information. The expressions "not" and "if . . . then . . ." are on Quine's list of logical constants, and those expressions occur in:

> (5) If Smith is at home, then it's not the case that Smith is not at home.

Before we proceed, we have to make a couple of points about the word "not." That word *not* is known as a "sentence-level operator": it's an expression that forms sentences out of sentences—it forms "snow is not white" out of "snow is white." It isn't often that sentence-level operators are replaced with variables. But not only *can* such replacements be made; they must be made. Otherwise, it would be impossible to express any generalization about sentence-level connectives. (Logicians and linguists often make statements about whole classes of sentence-level connectives, and such statements involve replacements of this kind, since they have the form: "given *any* sentence level connective *X* . . .") In fact, we're about to state just such a principle right now.

Given (5), if we replace "not" with a variable, the result is:

> (5_X) If Smith is at home, then it's *X* the case that Smith is *X* at home.

If we bind the free variable with a universal quantifier of the appropriate kind, the result is:

> (5_{XUG}) Given any sentence-level operator *X*, if Smith is at home, then it's *X* the case that Smith is *X* at home,

which is false, the reason being that some sentence-level operators lack the relevant property. An example of such an operator is the word "necessarily." When this word replaces the variable in (5_X), the result is:

(5_{XN}) If Smith is at home, then it's necessarily the case that Smith is necessarily at home,

which is false.

Another expression on Quine's list that occurs in (5) is "if . . . then . . ." This expression, like "not," is a sentence-level operator. But whereas "not" is a *one*-placed sentence-connective, "if . . . then . . ." is a *two-place* sentence connective. So, whereas "not" assigns sentences to single sentences, "if . . . then . . ." assigns sentences to ordered pairs of sentences. Given the ordered pair <Toby is a dog, Toby is a mammal>, "if . . . then . . ." yields the sentence: "if Toby is a dog, then Toby is a mammal."

Everything just said about "not" is true of it. If it's replaced with a variable, which is then bound by a universal quantifier of the appropriate kind, the result is:

(5_{IT}) For any two-place sentence-connective *X* <Smith is at home, it's not the case that Smith is not at home>

But (5_{IT}) is simply false. Let *C* be a two-place sentence-level operator defined as follows: Given two propositions (or sentences) *P* and *Q*, *C*(*P*,*Q*) is true exactly if *P* is incompatible with *Q*. In that case:

(5_{C}) *C*<Smith is at home, it's not the case that Smith is not at home>

is false.

What we just said about "not" and "if . . . then . . ." is true of all the other items on Quine's list of "logical constants." And it's true of many other expressions that Quine did not personally include on that list but that, according to many, nonetheless belong on it. (For example, some hold that "necessarily" and "possibly" belong on that list.) This suggests that, when Quine judged an expression to be a "logical constant," he did so on the (cogent) grounds that replacing occurrences of it with variables systematically failed to yield universally true open-sentences.

This suggests that the following definition of "logical constant" might be the right one (we'll soon see that it's only approximately right):

(LC[234]) For an expression *E* to be a "logical constant" is for there to be some universally-true open sentence that contains *E*.

LC is close to the truth, but it falls just short. Consider the open-sentence:

(#) If *P* and *Q* and grass is green, then grass is green.

(#) is universally true. But, since "grass is green" isn't a logical constant (#) is false. But, apart from "*P*," "*Q*," and "*R*," all the expressions occurring in:

(##) If *P* and *Q* and *R*, then *R*.

are logical constants. (##) is universally true, and it contains no expressions that, when replaced with free-variables, result in some *other* open sentence that is universally true. If a universally true open sentence contains no such variables, let us say that it is "pure." This shows that a "logical constant" is any expression other than a variable that occurs in a pure, universally true open-sentence.

7.5 analytic truth ≠ formal truth

As we discussed in Chapter 1, the two main contentions of Wittgenstein's *Tractatus Logico-philosophicus* (TLP) are (i) that all non-empirical truths are tautologous (i.e., that they are definitional truths; e.g., "a fortnight is a period of two weeks") and (ii) that all non-tautologous truths are observation-reports.

In Chapter 1, we saw why each contention is false. Given the points just made, we are in a position to put forth another reason to reject (i) and, therewith, to reject a third contention of the TLP, viz., that all analytic truth is formal truth and, therefore, that all entailment is formal entailment.

The open-sentence "If *P* and *Q* and *R*, then *R*" is true for all values of its variables. This means that:

> (A) For any propositions *P*, *Q*, and *R*, ‹if *P* and *Q* and *R*, then *R*›[235] is true.

(A) is true, of course. But it is not formal truth. For there are false sentences that have the same form. Here is one such sentence:

> (B) For no propositions *P*, *Q*, and *R*, ‹if *P* and *Q* and *R*, then *R*› is true,

In the place where (B) contains an occurrence of one expression ("for *no* propositions *P*, *Q*, and *R*"), (A) contains an expression with a different *meaning* ("for *any* propositions *P*, *Q*, and *R*"). So (A) and (B) are no more *formally* different than "Jim is tall" and "Mary is tall." Thus, the fact that A isanalytic, whereas B is not, cannot possibly be chalked up to the fact that one is formally true, whereas the other is not. Thus, *not all analytic truth is formal truth.*

In the TLP, Wittgenstein[236] says that all analytic truth *is* formal truth. Wittgenstein is wrong, for the reasons just given. Also, even if there were some formally true sentence *S* such that *S* was equivalent with (A), the statement (in this context you make take "statement" to mean "sentence," "sentence-token," or "proposition"):

> (C) *S* is equivalent with A.

would not *itself* be a formal truth. To be a formal truth is to be a consequence of an informally true universal generalization. The sentence:

> (%) if grass is green and snow is white and roses are red, then roses are red

is a formal truth because it is a consequence of (A). Thus, Wittgenstein's thesis that all analytic truth is formal truth is not only wrong but incoherent.

Here is another illustration of that fact. The sentence:

(JTB) any case of knowledge is a case of justified true belief,

is an analytic, not an empirical, truth. But it isn't a formal truth, since it has the same form as:

(NTB) any case of knowledge is a case of unjustified false belief,

which is false.

Wittgenstein grants that there *seem* to be analytically true sentences that aren't formally true. But he denies that there really are such sentences. His position is that, although it *seems* not to be formally true, the sentence:

> (JTB) any case of knowledge is a case of justified true belief,

is in fact formally true, the reason being that it is an abbreviation of:

> (JTBF) any case of non-accidentally true, justified belief is a case of justified true belief,

or, at any rate, of some other formally true sentence.

Contrary to what Wittgenstein alleges, JTB is not JTBF in disguise. They express different propositions. JTB is informative. JTBF is not. One can accept the proposition meant by the one without accepting the proposition meant by the other. JTB and JTBF express *equivalent* propositions (assuming, as we are, that JTBF is correct). But if JTB is to be a mere

abbreviation of JTBF, both must express the very same proposition. They don’t; so the view that analytic truth is the same thing as formal truth (henceforth “AFT”) is false.

There is another problem with AFT. Any given informal analytic truth can be turned into a formal truth by replacing expressions with analytic definitions of themselves—by replacing “circle” with “closed planar figure of uniform curvature,” "2" with "1+1", and so on.

But there's a deeper problem with Wittgenstein's contention that:

> (AAT)[237] All analytic truths are either formally true or they are abbreviations of such truths.

AAT is to the effect it is itself false. Obviously AAT isn't a linguistic convention; it isn't like "the term 'yard' refers to a distance of 3 ft" Nor is it formally true. (It has the same form as "all cats are either birds or big birds," which, being false, isn't a truth of any kind.) Nor, obviously, is AAT an abbreviation of a formal truth. For, if it *were* an abbreviation of such a truth, it would, given it itself says, be an abbreviation of:

> (AATF) All formal truths are either formal truths or abbreviations of such truths,

which, being trivial, cannot be what AAT means.

Since it is neither formally nor conventionally true, one of the following statements must be true:

> (i) AAT is false
> (ii) AAT is an informal, analytic truth
> (iii) AAT is an empirical truth.

To accept (i) or (ii) is to reject AAT. So an advocate of AAT must hold that it's empirically true, i.e. that, bur for the observations we've had, we couldn't possibly know it to be true and, therefore, that it's negation is, from a logical viewpoint, no less likely to be true than AAT itself.

But it cannot possibly be an empirical question whether there are informal analytic truths. An analytic truth is one that cannot possibly fail to hold. So if there are *informal* analytic truths, they hold in every possible circumstance:

given a statement P that is either analytically true or analytically false, there is nothing about the spatiotemporal world, and therefore nothing one could learn through observation, that would establish or even confirm P, supposing it true, or not-P, supposing it false.

7.6 Conceptual analyses never expressed by formal truths

First a terminological point: The sentence "condition *C* is fulfilled" is synonymous with "property *P* is instantiated."

A ***conceptual analysis*** is what is expressed by a non-empirical, non-trivial sentence of the form "C_1 is fulfilled if and only if C_2" where C_1 and C_2 are conditions. (To save time, I'll sometimes speak as though conceptual analyses are *identical* with sentences.) Thus:

> (CK) *x* is a circle if and only if *x* is a closed planar figure of uniform curvature,

is a conceptual analysis, and so is:

> (CK*) *x* is a closed planar figure of uniform curvature iff , for some plane *P* and some point D on *P*, *x* is the area bounded by every point on *P* that is a fixed distance from *D*.

Neither CK nor CK* is a formal truth.

No conceptual analysis is formally true. Conceptual analyses are expressed by analytic statements, and analytic statements either: (i) coincide with informal, analytic universal generalizations or (ii) are instances of such generalizations. If (i), they're obviously not formally true. But any statement that is an instance of some informal, analytic universal generalization is trivial, and is therefore not analytic. For example, % is a consequence of (A), and % is clearly trivial.

Part III Knowledge and Inductive Inference

Chapter 10 Knowledge

1.0 The ambiguity of the word “know”

The term “know” is ambiguous. Sometimes it refers to awareness of *objects,* as in “I know Smith.” But oftentimes it refers to awareness of *truths,* as in “I know that snow is white.”

Other languages use different verbs to mark this difference. For example, in Spanish, when one speaks of knowing *people,* one uses the verb “conocer,” and when one speaks of knowing truths, one uses the verb “saber.” German also uses different words to mark this distinction (“kennen,” for knowing people, and “wissen” for knowing truths). Russian does not. It is like English, and uses one verb—“znat”—for awareness of both people and truths.

This last fact is suggestive. If English were the *only* language that had just one word for both kinds of awareness, its doing so would be a mere idiosyncrasy—a random fact about one language that didn’t reveal anything inherent in human thought. But the fact that *other* languages do the same thing suggests that there is a general tendency *not* to distinguish these two very different kinds of knowledge. We will find that, in fact, this tendency of ours has a solid epistemological basis and is therefore not a logical *faux pas.* However counterintuitive it may seem, knowledge of people *is,* we will find, knowledge of truths.

1.1 The philosophical meaning of the word “know”

In philosophy, when we talk about “knowledge,” we are primarily, but not exclusively, concerned with knowledge of *truths.* Nonetheless, we will find that the distinction between knowledge of truths, on the one hand, and knowledge of individuals, on the other, is of great importance.

We will find later that the word “know” is actually ambiguous in additional ways. For example, psychologists and philosophers often talk about “procedural” knowledge. Any case of anyone’s being able to do anything is a case of procedural knowledge. My ability to play the piano is a case of procedural knowledge; so is my ability to speak English; and so is my ability to tie my shoes.

Many deny that procedural knowledge is *bona fide*knowledge. This is the exact opposite of the truth. What the virtuoso knows procedurally, the incompetent neophyte knows only discursively. As it deepens, knowledge becomes increasingly proceduralized. Non-procedural knowledge is poorly integrated knowledge; it is knowledge that hasn’t yet established the links to other pieces of knowledge that are needed to mobilize it. Procedural knowledge is knowledge that is sufficiently well-integrated to have established these links. It is therefore indefensible to hold that procedural knowledge is non-knowledge. We’ll say more about this in Section 9.3. But for the time being, when the word “know” is used, assume that it is referring to knowledge of truths, unless there is an explicit indication to the contrary.

It should be pointed out that, even though, for the reasons given, the word “know” *seems* to have many meanings, some philosophers maintain that ultimately all knowledge is identical with knowledge of truths—in other words, that awareness of objects and also procedural knowledge, along with the other kinds of knowledge that we will talk about, are ultimately just different ways of knowing truths. But this is extremely controversial.

2.0 Knowledge as justified true belief

One cannot know something false. Of course, people often firmly *believe* that they know things that are in fact false. People can believe that the Earth is flat with as much conviction as they can believe that the Earth is round. The distinction between a true belief and a false one is not always marked by internal psychological or emotional differences. For example, one cannot tell whether a belief that one has is true or false merely on the basis of the intensity of that belief. In any case, it is a simple fact that only truths can be known.

All cases of knowledge are cases of belief. If you know something, you believe it. If you know that Paris is the Capital of France, you believe it. If, speaking non-ironically, somebody said "I know that Paris is the capital of France, but I don't believe it," that person would be uttering nonsense.

But not all instances of belief are instances of knowledge. We've already seen one reason for this. False beliefs are not knowledge. But there is another reason why not all belief is knowledge. A belief must be *justified* if it is to be knowledge.

A lucky guess is not knowledge. Suppose that you're taking a multiple-choice test and you come across the following question:

> With what digit does Professor Kuczynski's phone number end?
>
> (a) 9
> (b) 3
> (c) 2
> (d) 0
> (e) none of the above

A lucky guess will be enough for credit on the test, but it won't be enough for knowledge. You must have some good reason for your choice. Your opinion as to what the right answer is must be *justified* for it to qualify as knowledge.

There are a lot of ways that you can have such a justification. Maybe I've told you my number. Maybe you saw it on one of my class handouts. Maybe you hired a private detective to find it out. All of these would be good

sources of justification. But without some such source, you don't have knowledge.

Also, spurious reasons that happen to lead to the right answer are not enough for knowledge. Suppose that you believe, wrongly, that all people with red hair have cancer; and also suppose that, for that reason, you believe that your red-haired friend Bob has cancer. It may be that Bob *does* have cancer. But your reason for believing this is spurious; so you don't *know* it.

False premises can lead to true conclusions; misinformation can lead to correct beliefs. Of course, accurate information is much more *reliable* than misinformation. A person whose methods are sound and whose premises are correct is considerably more likely to find out the truth than he would be otherwise. But false premises and unsound methodologies can lead to truth. When it happens, it's a kind of accident, as our previous example shows. But accidents do happen.

So for something to be knowledge it must have the following three characteristics:

(1) it must be true;
(2) it must be a belief; and
(3) it must be justified.

Thus, all cases of knowledge are cases of justified true belief.

2.1 What constitutes justification?

We must define a few more terms before we can answer this question. The term “perceptual” means having to do with sense-perception—that is, with sight, hearing, sound, taste, smell, bodily sensations (e.g., stomach aches), and also sensations of motion and resistance. I know that there is a computer in front of me because I can see it. (In any case, this is part of the story: there’s more to it, as we’ll see later.) I know that the stove is hot because I touched it. I know that one cup is heavier than another because I feel a greater degree of resistance when I try to move the one than when I try to move the other. (Notice that this last kind of perception isn’t identical with one’s sense of touch, even though it involves the use of touch.)

The term “introspective” refers to the direct awareness that one often has of one’s own mental contents. Right now I am aware of the fact that I am having thoughts—that, more specifically, I am having thoughts about the theory of knowledge (*I’m writing about it*), about this chapter (*I hope that I’ve finished writing it by tomorrow*), and about tonight’s dinner (*it’ll be chicken*). My knowledge of this is not based on the use of my senses: I don’t *see* or *touch* those thoughts. At the same time, that knowledge seems not to involve my making inferences: I seem to be directly, non-inferentially aware that I am having them.

Bearing all of this in mind, let us consider the question: under what circumstances is one justified in having a belief? Many different kinds of facts can be known; and, as we’ll see, the way in which a belief must be justified for it to be knowledge depends on the category into which the fact in question falls. But there are ultimately only three sources of justification: (a) perceptual evidence, (b) introspective evidence, and (c) rational inference. Some cases of knowledge involve *only* (c); some cases of knowledge are a combination of (a) and (b); the majority of cases involve a combination of both (a) and (c). Some philosophers[238] have argued that whenever (a) is involved, so is (b).

2.2 Can testimony constitute a source of justification?

There is much that a given person knows that he hasn't personally verified—much that he knows because he is *told* it. Such telling can take various forms: it can consist of written or spoken speech, photographs, and drawings. (This list is not complete.)

I haven't personally looked at the Earth from outer space. But I still *know* that it is approximately spherical. How do I know this? Because I've been told it. Using speech, writing, photographs, drawings, and possibly other media, various people have told me various things that justify my belief that the Earth is spherical. For similar reasons, I have at least some knowledge about what kinds of activities are likely to bring about lung cancer and about the political situation in distant parts of the Earth, even though I have not personally done cancer research or personally visited those parts of the Earth.

But, of course, not *all* beliefs that are based on testimony qualify as knowledge. First of all, much of what we are told is false. But even true beliefs that are based on testimony are not always knowledge. To qualify as knowledge, testimony-based belief must (a) come from a reliable source and (b) the person being given the testimony must have independent evidence of the reliability of that source.

A story may clarify this. I belong to a cult. My cult leader wants me to drink a certain beverage. He has told me that it is extremely nutritious. As it happens, it *is* nutritious—the cult leader has told me the truth. Further, I *believe* that the beverage is nutritious. But the reason for this is *not* that I have any actual evidence that the cult leader is well-informed or truthful; the reason is that, for purely emotional reasons, I believe everything the cult leader tells me.

Here I have a true belief—I believe, correctly, that the beverage is nutritious. But do I *know* this fact? No. And the reason is that I don't have good reason to believe that the cult leader is in fact telling the truth—even though, in this case, he is telling the truth. The data that I have concerning the cult leader—more specifically, concerning the likelihood of his stating truths as opposed to falsehoods—does not warrant my accepting what he is saying.

Here's another, similarly-themed story. A world-renowned physician has told me that I have arthritis. His opinion is based on his expertise and also on the results of various tests that he has performed on me. He is right—I do in fact have a mild case of arthritis. Further, I am aware of the fact that he is an

arthritis expert and that he has conducted various tests.

In this case, I obviously *do* have reason to believe the testimony I am being given. My belief is therefore justified and it is, or at least could be, a case of knowledge. (I say “or at least could be” because, for reasons that we’ll discuss later, even beliefs that are true and justified aren’t always knowledge.)

So a testimony-based, true belief can be knowledge. But it is necessary that the person who has that belief have some reason to believe that the source of the testimony is accurate.

Notice that testimony-based belief involves a combination of perceptual knowledge and also of rational inference. You must perceive the words that are being told to you. (You hear them or see them, depending on whether they are written or spoken.) You must also have perceived some kind of evidence of the truthfulness of the source of testimony. Finally, from those two sources of perceptual information, you must infer that the source of testimony is reliable. So testimony-based knowledge involves a combination of (a) and (c).

2.2.1 A point concerning testimonial knowledge

When we're growing up, we're told many true things by parents, teachers, and other figures of authority. During our earlier years, we obviously don't have independent corroboration for the testimony of these people. And yet we're inclined to say that we *know* the things they told us.

What I believe to be going on is that, at first, we *didn't* know these things. We just believed them. But then we either came to have evidence of the trustworthiness of the people who told them to us or we discovered them for ourselves.

3.0 Knowledge as hyper-justified true belief

Not all cases of justified true belief are cases of knowledge. It is only when one has a belief that is true and that is justified in a rather specific way that one has knowledge. More specifically, one has knowledge only when one has a belief that is true and justified and, moreover, *one's being justified in holding that belief is non-accidentally related to that belief's being true*.

A story will help clarify this. You work at some company. Smith is a fellow employee of yours. He is an excellent worker; the boss likes him; he's overdue for a promotion; and the boss, who is honest and rewards good work, has just announced that he's going to give a promotion to somebody (but he hasn't said who). Further, you know all of these facts. You also know that Smith drives a Ford. On the basis of this information, you form the belief that:

> (*) Somebody who drives a Ford is about to be given a promotion.

Obviously this belief is justified, since your reasons for holding it are solid.

And that belief turns out to be true—but not for the reason that you have in mind. Smith does *not* get the promotion. The boss gives the promotion to Jones. Jones is a terrible employee, as you know. But what you *don't* know, and had no way of knowing, is that the boss owes a huge favor to Jones' dad, who is a local and exceptionally vindictive Mafia boss. So, acting with an uncharacteristic disregard for merit, the boss gives the promotion to Jones. By sheer coincidence, Jones drives a Ford.

In light of this, let's look at your belief in (*). First of all, that belief is true—a Ford driver (namely, Jones) *was* given a promotion. Second, it's justified. We've already seen why. You were justified in believing that Smith, whom you knew to be a Ford driver, was going to be a given a promotion. That belief turned out to be wrong—but it was still justified.

So your belief in (*) is true and justified. It is thus a justified true belief.

But is it knowledge? No. It is pure coincidence that Jones, who you had no way of knowing would get the promotion, drives a Ford.[239] You're right a Ford driver would get the promotion, and your belief is justified. But you still believe it for the wrong reasons. (A justified belief can thus be had for the

wrong reasons.) And that's why it isn't knowledge.

This shows that there can be justified true beliefs that are not knowledge. It shows that for something to be knowledge it is not *sufficient* that it satisfy conditions (1)–(3), even though, for the reasons discussed earlier, it is *necessary*.

The argument just stated was given by Edmund Gettier (1963). Bertrand Russell (1912) gave a similar argument. You look at a clock that is in fact broken but that you believe, and have every reason to believe, is in good working order. According to the clock, the time is 4:00 P.M. In fact, that *is* the actual time. But the clock is broken, and it is therefore a fluke that it gave you the right time. Your belief that it is 4:00 P.M. is thus true and justified, but it isn't knowledge.

3.1 Knowledge as non-accidentally justified true belief

We've seen that for something to be knowledge, it isn't enough that it be a justified, true belief, and that it must therefore satisfy some fourth condition. There is considerable debate as to the precise nature of that fourth condition, but it seems clear that, when stated very approximately, it must be this:

> (4) if a given belief is to be knowledge, there must be a non-accidental connection between one's reason for having that belief, on the one hand; and the fact in virtue of which that belief is true, on the other: justification and truth must be non-accidentally connected.

So for something to be knowledge is for it to satisfy the following *four* conditions:

> (1) it must be true;
> (2) it must be a belief;
> (3) it must be justified; and
> (4) there must be a *non-accidental* connection between the justification for the belief and the fact which makes the belief true.

Thus:

> (K) knowledge is *non-accidentally* justified true belief.

In other words, there must be some kind of *principled*, as opposed to a strictly fortuitous, relationship between the justification of the belief on the one hand, and its being true on the other.

But K is much too vague to be considered a solution to the problem at hand. What does the word "non-accidentally" mean in this context? Until we know the answer to that, we don't know what knowledge is. We'll now try to supply that answer.

3.2 This position refined

Let's consider Russell's clock story. You were right to believe that it's 4:00 P.M., and you are also justified in having that belief. Your immediate justification consists in your belief that:

> (RT) The clock is now telling me that it's 4:00 P.M.

But *by itself* a belief in RT wouldn't be enough to justify you in believing that it's 4:00 P.M. If you had no grounds for accepting RT, or only bad grounds, your acceptance of it would not provide you with any justification for believing the time to be 4:00 P.M. But you are justified in accepting RT; and here is that justification:

> (OB) In my extensive past experience, the clock has always been right. I also know that the building superintendent, who is in charge of making sure all the clocks are working, is very capable and diligent. These facts ensured that, if the clock says that it's a certain time of day, that's because it really is that time of day. That's why *the clock is now telling me that it's 4:00 P.M.*

Of course, the italicized part of OB is identical with RT.

Notice that OB is an explanation of a fact—of the fact that the clock is now telling you that it's 4:00 P.M. It is on the basis of your acceptance of RT that you believe the time to be 4:00 P.M. But OB is the wrong explanation of RT. The right explanation is that the clock is broken and, for that reason, always reads 4:00 P.M.

A story will help us move forward. You believe that Larry is in Texas, and your belief is justified. Earlier today, you saw Larry get on a plane. By all indications, the plane was Texas-bound. (It's flight 71 and, according to the overhead monitor, that flight is bound for Texas.) When Larry got to Texas, he called you on his cell phone/video phone. Somebody who looked exactly like him, and therefore probably *was* him, was standing next to a well-known Texas landmark. On that basis, you believe that, indeed, Larry is in Texas. And, in fact, he *is* in Texas.

Moreover, no off-the-wall, Gettier-type event infiltrated this sequence. In

other words, what happened is nothing like the following:

> Alternative Scenario (AS): Larry got on the plane. But Martians re-routed the plane to Barbados. For some reason, they then transported Larry to Texas. In Texas, somebody other than Larry, who looks and sounds just like him, calls you and tells you that "he" (which you take to mean "Larry") safely arrived in Texas. On that basis, you believe that Larry is in Texas. Which he is, and which you're justified in believing, but which, clearly, you don't know to be the case.

No—neither AS nor anything at all like it occurred. What happened is exactly what you thought happened. And you therefore know that Larry is in Texas. (Let's say that this story—the one where you *know* that Larry is in Texas—takes place in the "real world," even though, of course, it's no less a hypothetical than AS.)

Why do you know it in the real world, but not in the world described by AS? Let's look at the data on which, in the real world, your belief is based. Although you see Larry get on the plane, you don't see him deplane in Texas. You *do* see various images (on your cell phone/video phone) that are consistent with his being in Texas; and you encounter a lot of other things that, were Larry not to be in Texas, would be very hard to explain. Your belief that Larry is in Texas (after sitting on a certain plane, etc.) is not only justified: it also constitutes a correct explanation of the data that justifies your belief that he is in Texas. Your belief that, after sitting on a certain plane (etc.), Larry is in Texas, correctly explains *why* your cell phone is giving rise to these various images and noises.

Thus, your belief is not only justified, but *hyper*-justified. A belief is hyper-justified if it's justified and, moreover, embodies a correct explanation of the data that is doing the justifying.

In AS, your belief is justified, but not hyper-justified. Your belief is based on the same images and, in general, the same data that it's based on in the real world. But, in AS, your beliefs as to what caused those images and noises to occur are wrong. (You explained their occurrence by wrongly supposing that Larry produced them after being transported by non-Martians to Texas.) So, in AS your justified belief isn't knowledge because it isn't hyper-justified. I propose, then, that:

(KN) Knowledge is hyper-justified true belief.

4.0 Conceptual (analytic) vs. empirical (synthetic) knowledge

Some knowledge is purely conceptual; and some knowledge has a non-conceptual component. Knowledge that is purely conceptual is described as "analytic." Knowledge that has a non-conceptual component is described as "synthetic."

Consider the statement:

(#) Circles are figures of uniform curvature.

is true. And you know that it's true. You thus have knowledge of #. How does one become justified in believing #? Answer: by examining the statement itself. More specifically, by examining the components of that statement—the concepts *circle, curvature, uniform,* and so on. So your knowledge of # is purely conceptual. In other words, it is "analytic."

Analytic knowledge is so-called because it is acquired by *analyzing* the object of belief. Whenever statement-analysis is sufficient, by itself, to yield knowledge, that knowledge is analytic.

Most knowledge is non-analytic. Consider the statement:

(##) Donald Rumsfeld was born in the United States.

An examination of that statement is not enough to know whether it is true or false. It doesn't matter how much you dissect it and examine its components. In order to be justified in having a view, one way or the other, as to whether it is true, and in order therefore to *know* whether it is true, it is necessary to consider matters of fact of which no amount of conceptual analysis will apprise you. It is necessary to look at birth records or to ask informed people, or some such.

Thus, if one knows that ## is true, that knowledge is *non*-conceptual. In other words, it is *non*-analytic or, as philosophers sometimes say, *synthetic.*

4.1 More on the analytic-synthetic distinction

Although some philosophers have held otherwise, I think it is fair to say that synthetic knowledge is identical with knowledge of spatiotemporal facts, whereas analytic knowledge is knowledge of properties.[240] To know ##, for example, is to know something about the spatiotemporal world—the world of things that occupy space and time. By contrast, to know # is to know about unchanging, non-temporal (and therefore non-spatial) relations among properties.

In this chapter, we will operate on the reasonable, though debatable, assumption that all synthetic knowledge is knowledge of some fact or other about the spatiotemporal realm. So, in this context, the terms “spatiotemporal knowledge” and “synthetic knowledge” may be used interchangeably.

5.0 Knowledge of specific facts vs. knowledge of dependence-relations among facts

There is an important division within the category of spatiotemporal, or synthetic, knowledge. There is knowledge of *specific* facts, and there is also knowledge of *causal* or *nomic* relations among facts.

The term "nomic" means "having to do with natural law." Consider the fact that there is a combustion whenever a lit match is brought in contact with gasoline. Obviously that is not a coincidence. There is clearly some kind of principled relationship between, on the one hand, something's being gasoline that is brought into contact with flame and, on the other hand, something's combusting. Some kind of *causal* relationship is at work.

Technically speaking, not all nomic relations are causal. But the distinction between causal and nomic relationships is a rather subtle one, and in this chapter we will ignore it. (In Chapter 16, it is made clear why not all nomic dependencies are causal dependencies.) Causal relations are instances of *natural laws.* If *x* causes *y*, that is because there is some law of nature to the effect that things that, in some relevant respect, resemble *x* always bring about things that, in some relevant respect, are like *y*.

Causal relations typically involve *sequences* of events. But not all sequences of events are causal. When I bring my hands together quickly, a noise occurs—there is a clapping sound. That sequence of events is clearly causal in nature: the collision of my hands caused the sound to occur. But suppose that, just after I bring my hands together, one of you suddenly feels a tickle. That is not a causal sequence: it's just a coincidence.

The natural law of which a given causal sequence is an instance, is known as the *covering law* for that sequence.

To know a natural law is to know a fact about the spatiotemporal world. Natural laws are not in the same category as #. Natural laws are not of a strictly logical, philosophical, or mathematical nature. It is not possible to know of natural laws merely on the basis of conceptual analysis. Such knowledge can be acquired only through investigation of the spatiotemporal world. It is only through such investigation that one can know that the speed of light is 186,000 mph/second, that objects gain mass as they accelerate, and that the force with which two bodies attract each other is proportional to their masses.

Of course, some natural laws strike us as extremely obvious. We find it

obvious that unsupported objects fall. But that is not because such knowledge is analytic; it is only because the relevant spatiotemporal knowledge constantly forces itself upon us.

Even though knowledge of natural law is spatiotemporal knowledge, such knowledge tends to have a highly *general* character and is thus, in that respect, comparable to many cases of analytic knowledge (e.g., one's knowledge of #). Knowledge of the laws of physics is very general knowledge. To know such laws is not to know this or that specific fact, and is instead to know a fact about how events *in general* are interrelated.

But most spatiotemporal knowledge is highly specific. You know where you live; you know what toothpaste tastes like; you know where you parked your car; you know that Barack Obama is now the president. Such knowledge concerns matters of specific fact and not dependence-relations holding among facts in general.

It must be emphasized, however, that even though knowledge of natural law is general and is, from that viewpoint, similar to knowledge of logical or philosophical principles, knowledge of natural law is synthetic, whereas knowledge of logico-philosophical truths is analytic.

6.0 Knowledge of the future, the past, and the merely possible

Many future facts are not known and many future facts cannot *possibly* be known. But surely we are able to know at least *some* facts about the future.

One can know about the future because:

> (i) On the basis of sense-perception, one can know about the present,

and:

> (ii) One can know general principles, or causal laws, that link the present with the future.

Right now I know that there is water in the pitcher in front of me. I also know of the general fact that, under conditions relevantly like those that obtain in my kitchen, an open, overturned pitcher containing non-viscous liquid will release that liquid. Further, I know that I have an intention to tip the pitcher, so as to fill the glass in front of me with water. On the basis of these pieces of knowledge, I can obviously have a non-accidentally justified true belief to the effect that, in a moment, water will be flowing from the pitcher into the glass.

Knowledge of the future always has two ingredients: first, knowledge of present conditions, and second, knowledge of causal laws that, given circumstances similar to those known to obtain, yield certain outcomes. In some cases, the causal laws in question are of a relatively non-theoretical character. This was the case in the example, given a moment ago, concerning the water in the pitcher. In other cases, knowledge of the future may have an extremely theoretical character. This would be the case when, on the basis of relativity theory, a physicist predicts that the mass of a certain body will increase by a certain amount, by a certain time. But, in all cases, knowledge of the future is constituted by knowledge of causal laws[241] that link known current states of affairs to future states of affairs.

6.1 Knowledge of the past

We can obviously have knowledge of the past. There are two forms that such knowledge takes. In some cases, knowledge of the past is *direct:* one actually *remembers* the fact in question. I remember driving to work yesterday. My knowledge of this fact doesn't involve inferences made on the basis of causal laws.

But in other cases one's knowledge of the past *is* indirect, and all such knowledge *does* involve some knowledge of some kind of causal law. We know facts about the political institutions of ancient Rome. We obviously don't remember such facts, since we didn't personally experience them. So how do we know them? We have knowledge of certain current facts (e.g., certain currently existing records and ruins); and we *also* have knowledge of causal principles on the basis of which, given our knowledge of those current facts, inferences concerning the past—specifically, concerning what went on in Rome thousands of years ago—can be made.

Here's a simple example of the principle in question. You know that if you press an ink-filled pen against paper, inky lines will be deposited on the paper. (You also know, so let us suppose, that such lines are not likely to be deposited on paper in any other way.) This is knowledge of a causal principle. On the basis of your knowledge of this principle, plus your knowledge of current conditions, you can acquire *either* knowledge of the future *or* knowledge of the past. If you see that Smith is about to press his pen against the paper in his notebook, you know a fact about the future—you know that inky lines will soon result. On the other hand, if you see inky lines that were clearly made by a person, you know a fact about the past—you know that somebody dragged a pen across a certain piece of paper. Our knowledge of past events that we don't personally remember (e.g., facts about ancient Rome) falls into this second category.

When it is indirect, knowledge of the past has a structure that is similar to knowledge of the future. More specifically, past knowledge can be thought of as *inverted* future knowledge. We know the future when, using known causal laws, we project forward from known current conditions. We know the past when, using known causal laws, we project backwards from such conditions.

6.2 Counterfactual knowledge

We can have knowledge of what *would have been* (but wasn't). Even though Smith did not in fact fall off the ledge, I know that:

> (S) *If* Smith had fallen off the ledge, he would now be dead or severely injured.

Propositions like *S* are called *counterfactuals.* In general, a counterfactual statement is one of the form *if P, then Q*, where it is assumed that *P* is false.

Knowledge of counterfactuals is identical with knowledge of relations of either logical or causal dependence. To know *S* is to know of some causal law that, given a case of somebody's falling from a great height, leads to that person's breaking his or her leg. To know that ‹if *x* were a square, *x* would have four sides› is to know that ‹*x* has four sides› is a logical consequence of ‹*x* is a square.› (The symbols "‹" and "›" are called "quasi-quotation marks." For an explanation of what they are, see

Many contemporary philosophers take counterfactual statements to be statements about events in *other universes.*[242] So they take:

> (K) If Kennedy hadn't been assassinated on Nov. 22, 1963, he would have been reelected in 1964

to mean that:

> (K#) There is some other universe where Kennedy was *not* assassinated on Nov. 22, 1963, but, up until that time, was as much like our world as that fact allows it to be. In that universe Kennedy was reelected in 1964.

And they argue that:

> (S) "If Smith, who is six-feet tall, were half his current height, he'd be three-feet tall,"

means that:

(S#) There is some alternative universe where Smith is half his current height. In that universe, Smith is three-feet tall.

This whole line of thought is preposterous. Any knowledge that we have about other universes is parasitic on our knowledge of modal truths. First, we know the counterfactuals. Only then do we know what's going on in other worlds. It's only thanks to our knowledge that 6 divided by 2 is 3, along with our knowledge that Smith is six feet tall, that we know that Smith would be three feet tall in a world where he were half his actual height. It's only thanks to our knowledge that Kennedy was an enormously popular first-term President, along with our knowledge that popular incumbent Presidents win re-election, that we know that Kennedy would have won in 1964 had his life not been cut short.

These pedestrian points show that other worlds are useless in the way of giving us modal knowledge. It also shows that modal truths don't hold in virtue of what goes on in other worlds. *S* is obviously correct. Whatever it is that makes it correct, we don't have to leave this world to have access to it. It follows that, given any true counterfactual statement *S*, it is in virtue of what happens in this world, and this world alone, that *S* is true.

The real meanings of *S* and *K* are:

(S*) The statement that Smith is six-feet tall entails that the statement, for all values of *x*, ‹*x* is half Smith's height› entails ‹*x* is three-feet tall.›

and:

(K*) "Kennedy's assassination caused an otherwise unthreatened Kennedy not to be reelected."

For further discussion of the nature of counterfactual truth, see Chapter 7.

7.0 Theoretical vs. observational knowledge

It seems that much of what we can know comes *directly* from sense-perception. I open my eyes in a well-lit room, and I can just *see* that there are chairs and people there.

But much knowledge is theoretical. We don't have direct knowledge of electrons or even of gravitational fields. We have direct knowledge of the *effects* of these things. (More accurately, we have *relatively* direct knowledge of such effects. We'll see in a moment that our knowledge of these effects is itself indirect, even though it's obviously less so than our knowledge of electrons, etc.) But in characterizing something directly perceived as the effect of "gravitation" or of the activity of "electrons," we are going well beyond what we can literally see (or hear or touch, etc.), and we are assuming the existence of things whose existence can be justified only by *theories* of some kind or other. A theory is an explanation of what is directly experienced in terms of what is not directly experienced and whose existence is therefore not directly known.[243]

Freud and others have posited the existence of mental activity that we are not directly aware of and of which we are therefore "unconscious." One obviously can't be directly aware of unconscious mental activity. (Any mental activity that one is directly aware of is *ipso facto* conscious.) Thus, any knowledge that we have of the unconscious is inferential. In addition, such knowledge is also *theoretical*. But here we must take a moment to clarify the distinction between inferential and theoretical knowledge.

A theory is an integrated set of statements that, if correct, describe some sector of reality that cannot be known strictly through sense-perception or introspection. All theoretical knowledge is inferential, but not all inferential knowledge is theoretical. If I see smoke, and I infer on that basis that there is fire, my inference isn't theoretical. It doesn't presuppose the truth of a theory. It assumes the truth of the statement "where there's smoke, there's fire," which merely describes an observable concomitance, and therefore doesn't qualify as a theory. But, supposing for argument's sake that there is unconscious mental activity and that we can have knowledge of it, the inferences that yield such knowledge *do* presuppose theoretical knowledge. If I infer from the fact that you are accident-prone that you have hostility towards yourself, I am obviously making several assumptions as to the nature of the psyche. Further, those assumptions don't merely register observable

concomitances, and they aren’t independent of one another. Thus, those assumptions jointly constitute a theory. It follows that any knowledge that we have of the unconscious is of a theoretical character. In Sections 8.2.-8.2.3 of the present chapter, we’ll say more about the distinction between theoretical and inferential knowledge.

7.1 Is all knowledge theoretical?

Some epistemologists and cognitive scientists maintain that even perceptual knowledge is theoretical. This position may seem strange at first, but there is at least one reason to think that it may be true.

Even though *you* don't have to do any work to generate your own sense-perceptions, your mind and nervous system certainly have to do a great deal of work. Consider the disturbances of your sensory surfaces that lead to your sense-perceptions. If the surface of a rock or a piece of wood were similarly disturbed, no sense-perception would result. What is the difference between you and a rock? Your mind can, whereas a rock cannot, *process* the relevant disturbances in the right way. It's quite extraordinary that your mind is able to do this; for all it is given are various disturbances that, in and of themselves, bear no resemblance (or only an extremely abstract resemblance) to the things off of which the relevant photons (or whatnot) are bouncing. Nonetheless, on the basis of this meager information, your mind is able to generate perceptual representations that accurately depict those objects. The gulf between input and output—between, on the one hand, the surface-stimulations on the basis of which your mind generates its perceptions and, on the other hand, those perceptions themselves—bears a striking resemblance to the gulf between, on the one hand, the data on the basis of which a theoretical physicist posits the existence of a swarm of a certain kind of sub-atomic particle and, on the other hand, that particle-swarm itself. In the case of the theoretical physicist, that gulf can be bridged only by a theory. This suggests that the same is true of the gulf between disturbances of your eyes and the perceptions in which those disturbances eventuate.[244]

7.2 *A priori* vs. *a posteriori* knowledge

Thus, one must have knowledge *in order* to acquire perceptual knowledge. Given that even a newborn can have sense-perceptions, it follows that knowledge of some kind is coded into the innate structures of our minds.

Such knowledge is referred to as a priori knowledge. Thus, knowledge is a priori if it is coded into the very structures of our minds and therefore pre-exists any perceptions that we have.

If knowledge is *not* a priori, then it is referred to as a posteriori. Thus, a posteriori knowledge is *acquired* knowledge, and a priori knowledge is knowledge that is within us, in some form, from the outset.

It is a matter of considerable debate whether there is a priori knowledge and, if so, precisely what it is that is known a priori. (My own strongly held view is that there is a priori knowledge.) Among those who believe that we have a priori knowledge are Plato, Leibniz, Kant, and Chomsky. Among those who deny it are Aristotle, Locke, and Quine.

7.3 A second reason to think that perceptual knowledge is theoretical

Suppose that *X* is some stop sign. If you look at *X* from a certain distance —from, say, a hundred feet away—it will appear a certain way to you. Given only how *X* appears to you from that distance, you cannot tell whether you are looking at an object that is around eight-feet tall and a hundred feet away *or* an object that is 1,000-feet tall and much further away *or* an object that is six inches tall and very nearby. Nonetheless, you know, when you are looking at *X*, that it is not 1,000-feet tall and far away or six-inches tall and very close, and that it *is* around eight-feet tall and about a hundred feet away. Given that *X*'s appearance, by itself, does not reveal this to you, how do you know it? It is your background knowledge that, in conjunction with how *X* appears to you under the circumstances described, enables you to assign the right height to *X*. A continuation of our story may clarify this. Your friend Larry is right next to *X*. You can see the *comparative* heights of *X* and Larry; you can see that Larry is around three-fourths the height of *X*. So it is because of *X*'s appearance *plus* this background knowledge that you are able to determine *X*'s height.

This must be understood correctly. You can just *see* how tall, approximately, *X* is. But, for the reasons just given, it seems that background knowledge of the kind just described seems to be involved in, and to some degree to constitute, your seeing this.

An argument similar to that just given in connection with seeing *X*'s size can be given in connection with seeing any object's size. Indeed, similar arguments can be given in connection with seeing any fact about any object —in connection with seeing how fast an object is moving, what shape it has, how far it is from one, and so on. And extensions of those arguments suggest what is true of seeing facts about physical objects is true of *any* kind of perceptual awareness of *any* fact about them.

7.4. Inferential vs. theoretical knowledge

There is a reason to hold that not all knowledge is theoretical, namely that *many* inferences are involved in knowing a theory and, consequently, inferential knowledge is a precondition for theoretical knowledge. Some background points are needed to make it clear why, at least arguably, this holds.

First of all, not all inferences—not even all inferences that lead to correct conclusions—yield knowledge. If I infer that Bob has cancer on the grounds that he has red hair, my inference doesn't yield knowledge, even if, by coincidence, Bob does have cancer. The reason my inference doesn't yield knowledge is that the "inference rule" implicit in it—in other words, the principle that licenses it—isn't a good one. That inference rule is: "if a person has red hair, he or she has cancer." Since that rule obviously isn't "truth-conducive"—since, in other words, that rule isn't likely to yield correct conclusions, given correct premises—no inference made on the basis of it is legitimate. So for an inference that yields a correct conclusion to yield *knowledge,* one must make that inference on the basis of a legitimate inference rule.

Theories are *systems* of statements standing in various *dependence-relations* with respect to one another. (The word "statement" here is not meant to have any linguistic overtones and should be taken as shorthand for "truth or falsehood.") A dependence-relation holds between two statements if one of them supports the other—if, in other words, one of them provides some basis for inferring the other. Knowing that a dependence-relation holds between two statements involves either inferring the one from the other or, failing that, recognizing that such an inference could be legitimately made. For example, my recognizing that there is a dependence-relation between "there is fire at time *t* in place *p*" and "there is smoke at *t** in *p**," where *t** and *p** are neighbors of *t* and *p,* involves *either* my knowing one of them on the basis of the other *or* it involves my seeing that such inference would be legitimate. In either case, I have inferential knowledge. (Even if I happened to see the fire and the smoke, and thus didn't have to infer "fire at *t* in *p*" from "smoke at *t** in *p**" or *vice versa,* my knowing that a dependence-relation holds between them involves my knowing that the one could be legitimately inferred from the other. And *that* knowledge, surely, deserves to be characterized as inferential.) Since, as we noted a moment ago, knowledge

of dependence-relations precedes knowledge of theories, and since knowledge of such relations is inferential, it follows that inferential knowledge precedes theoretical knowledge.

Of course, those who deny that all knowledge is theoretical don't have this in mind. Usually those who hold that there is "direct" knowledge mean that there is no *inferential* knowledge—they use the terms "inferential" and "theoretical" interchangeably and therefore, since they're not interchangeable, loosely.

7.4.1 Inferential vs. theoretical knowledge (continued)

To be made cogent, the foregoing discussion must be qualified in a couple of ways. These qualifications will bring out some principles whose importance extends far beyond the points just made.

I said that a dependence-relation holds between two statements just in case one of them can be legitimately inferred from the other. Some might object to this on grounds like the following:

> Smith's crashing the car might have resulted from his being drunk. So a dependence-relation holds between the truths expressed by the statements "Smith was drunk at *t*" and "Smith crashed the car at *t**." Were it not for the first truth, the second wouldn't hold, and the second is therefore dependent on the first. But surely I cannot legitimately infer the second from the first. Given only that Smith is drunk, I can't infer that he'll crash his car. So, contrary to what you say, dependence-relations aren't always backed by legitimate inference rules.

First of all, it wasn't *merely* Smith's being drunk that led to his crash. It was his being drunk while being behind the wheel. In fact, more than this is involved. He wouldn't have crashed had the car in front of him not slowed down or if he hadn't been talking on his cell phone when this happened. It was a confluence of factors—his drunkenness, the behavior of the driver in front of him, his own distractedness—that led to the crash. Total causes are always confluences of circumstances. Thus, the dependence-relation in question holds, not between the two statements cited by the objector, but between "at *t*, Smith crashed" and "at immediately preceding time *t**, Smith was driving while drunk and talking on his cell phone, when the car in front of him suddenly slowed down." The second statement *does* make it reasonable to infer that, at or around *t*, Smith crashed.

Admittedly, the second statement doesn't make it *certain* that he crashed. This is because the presumption that, given the various factors at play (his drunkenness and distractedness, the sudden deceleration of the car in front of him, etc.) the imminent crash could still be headed off. Maybe the person in the passenger-seat of Smith's car wrenched the steering wheel to the right, just in time to avoid the crash. Not likely, but possible. Maybe Smith is

somebody whose driving is enhanced by drunkenness. Not likely, but possible. But even though, given the operative conditions, Smith's crashing cannot be known with certainty to occur, it doesn't follow that the inference from "Smith was drunk and chatting on his cell phone (etc.)" to "Smith crashed" isn't legitimate or, therefore, isn't licensed by a truth-conducive inference rule. What follows is that the inference rule in question is *defeasible.* An inference rule is defeasible if (i) it provides a reasonable *presumption* that, given certain premises, a certain conclusion holds, but (ii) that presumption may turn out to be false if other facts, not in evidence, happen to intercede.

There is a big difference between an inference rule that is defeasible, on the one hand, and one that is simply bad (i.e., not truth-conducive), on the other. A bad inference rule provides *no* reason for inferring one thing from another. A defeasible inference rule (when good) does provide such a reason. It doesn't provide as good a reason as a non-defeasible inference rule. But that doesn't mean that it's not good at all; goodness (truth-conduciveness) comes in degrees.

7.4.2 Rules vs. principles

There's also an important difference between *rules* and *principles*.[245] Rules are about outcomes. Principles are about tendencies. Rules have the form: "if this, then that." (If it's a raven, then it's black.) They're categorical. (All members of the category *raven* are members of the category *black*.) A rule that admits exceptions is false. (A single non-black raven falsifies the rule that all ravens are black.) A correct rule can be obtained by replacing a counter-exampled rule with one that is duly hedged with exception clauses. (So, supposing that Smith is the one non-black raven, then "all ravens other than Smith are black" is indeed a correct rule.)

But a duly hedged rule is not a principle. Principles are not watered down rules. It is no accident that so many drunk drivers crashed; it's not a statistical fluke. Nor is it parasitic on some other, more robust connection. In a society where the government forced blond people to live near nuclear waste, there would be a connection between being blond and getting cancer. Drunk driving is to crashing what being exposed to radiation is to getting cancer. It is not what, in the just-described society, being blond is to getting cancer.

As we'll see in Chapter 12, empiricists think of principles as being watered-down rules—as being expressed by categorical statements to which there are exceptions. This isn't the right way to think about them. Principles are better thought of as registering the existence of defeasible forces. Being drunk can be thought of as a car-crash-conducive force (or, more likely, confluence of forces) that, like most forces, can be overridden. Note that, whereas inference-rules are defeasible in a logical sense, forces are defeasible in a causal sense.[246]

8.0 Knowledge of ourselves and others

A person's knowledge of himself is obviously much better than his knowledge of others. You obviously have a much more accurate understanding of your own mind than you do of other people's minds. Nonetheless, we can, within limits, have some knowledge of what others are thinking.

It is a matter of considerable dispute how we can acquire such knowledge. But here is a hypothesis that, I think, must surely be correct—at least for some, though perhaps not all, instances of knowledge of others.

We can, at least within limits, have direct knowledge of what is going on within our own minds. And we often know how we are likely to behave in response to an experience of a given kind. When others behave in a given way, I tend to ascribe to them the kind of mental state that accompanies instances of *my* behaving in that way. I know that when I am in pain, I tend to wince or clutch at the wounded part of my body (or some such). So when I see somebody else behaving in that way, I tend to assume that he or she is in pain. Comparable points hold with respect to our ascriptions of other kinds of mental states to others.[247]

8.01 Wittgenstein's attempt to refute this argument

Wittgenstein (1958) rejects what we just said. He rejects the idea that one ever acquires knowledge of what is going on in another person's mind by assuming the latter's outward behavior to express a mental state similar the sort of mental state to which, in one's own case, such behavior gives rise. Wittgenstein argues that:

> (WA)[248] A given person has direct knowledge of only *one* case---namely, his own---in which mental states of such and such a kind are followed by physical behaviors of thus and such a kind. Were one to conclude solely on the basis of one's own case that, where bodies *other* than one's own were concerned, such and such mental states eventuated in thus and such behaviors, one would, in so doing, be drawing conclusions that vastly exceeded anything warranted by the available information.

WA is a text-book case of fallacious reasoning. In a celebrated passage, David Hume (1700-1776), though obviously not addressing WA specifically, made it clear why this is so.[249] Let us suppose that we discover some new kind of metal *M*, and we find that some one specimen *m* of *M* is conductive to a certain degree, malleable to a certain degree, has a certain micro-structure, and a certain melting point, etc. We know that, given any type of metal, its specimens are identical in respect of their microstructures, melting points, coefficients of expansion, etc. So if we assume, on the basis of what we know about *m*, that *all* specimens of *M* have such and such melting points, degrees of conductivity, etc., we are by no means over-drawing conclusions. For our inference is *not* based on *m* alone, but on *m along* with the previously mentioned background information. The other side of the coin is that, were *m* to *differ* in respect of its coefficient of expansion, etc. from other instances of *M*, then (supposing that we didn't decide on that basis to reclassify *m*) its thus differing would constitute a singularity of unprecedented proportions. It would thus be highly *counter*inductive to suppose that *m* might differ in these respects from other members of *M*.

It would be similarly counterinductive to suppose that it was only in one's

own case that certain anatomical and behavioral phenomena had certain mental accompaniments. What Wittgenstein describes as abstaining from overdrawing conclusions is in fact positing singularities for whose existence the past, if it's evidence of anything, provides strong *negative* evidence.

Also, to pick up on a parenthetical point made a moment ago, were one to discover that the bodies that one previously *thought* to be animated by minds like one's own were *not* thus animated, one would *ipso facto* reclassify oneself. One would *ipso facto* cease to regard oneself as being of the same ilk as those mindless entities, one's reasoning being: "they are zombies; I am not; so in at least one vitally important respect, I am not one of them."

Wittgenstein doesn't see this. He doesn't see that, *if* people other than oneself were to turn out to be zombies, they *ipso facto* wouldn't be "people", except in a narrowly biological sense. Wrongly assuming WA cogent, Wittgenstein believes there to be no choice but to deny the existence of minds. There is, he says, no "inner Cartesian theatre." What he means is that there is nothing about one's mind that is any more transparent to its owner than it is to any other person.

This idea is, quite obviously, false. People can conceal much about themselves from others. Others may never be able to know about a given person what he can immediately know about himself. But Wittgenstein will have none of this. "The best picture of the mind is the body," Wittgenstein famously says, meaning that there is nothing you know about yourself that I don't know by looking at you. "An inner state stands in need of an outward criterion," Wittgenstein says, meaning that the very idea of a mental state that doesn't involve overt behavior is absurd. So, if Wittgenstein is right, two soldiers, both of whom are standing perfectly still, cannot be in different psychological states.

No argument could make the spuriousness of Wittgenstein's position any more obvious than it already is. Nonetheless, in Chapter 22, we consider Wittgenstein's position and, for what little it's worth, put forth arguments that make it clear that it's wrong.

8.1 Self-knowledge

People know much about themselves. I know that, right now, I have the intention of producing a helpful essay on the theory of knowledge. I know that, right now, I'm in a pretty good mood. I know that I enjoy playing the piano. I know that I wish that I could play as well as Arthur Rubenstein; I also know that I'm sad that I cannot play as well he could. All of you have comparable knowledge about yourselves.

A striking feature of at least *some* of our self-knowledge is how direct it is. My knowledge that I'm in a pretty good mood is even more *direct* than my knowledge that I am now typing on a computer. But how is this possible? Isn't my knowledge that I'm now typing on a keyboard as direct as knowledge can be?

No. My knowledge that I'm typing on a keyboard is based on the sensory experiences I'm having. These experiences are both visual and tactile. ("Tactile" means "having to do with the sense of touch.") But, given only the nature of those experiences, it is a possibility, if only a theoretical one, that I'm dreaming or hallucinating.

My having accurate perceptions of the computer screen in front of me involves my body's being stimulated in very specific ways by causal processes that originate with that computer screen. (Light waves must bounce off of that computer screen and disturb my retinas.) Further, those light waves mustn't be distorted or altogether cut-off by intervening influences. If there were an uneven, thick piece of glass between me and the screen, those light waves would be distorted, and my perception of the screen would be inaccurate. If there were a piece of thick wood between me and the glass, those light-waves would be cut off, and I wouldn't see the screen at all. So my having an accurate perception of the screen involves more than just my having a certain visual experience: that experience must be caused in a very specific way by very specific events in the external world (i.e., the world *outside* of my body).

A corollary is that I am not in any sense seeing the screen unless I am having a visual experience that is caused by events involving the screen. Even if I'm having visual experiences that are subjectively just like those involved in my actually seeing the screen, I'm not seeing the screen *unless* those experiences result from disturbances of my ocular surfaces caused by their being struck by light-rays bouncing off of it. So I'm not seeing the

screen if light-rays responsible for those experiences are making their way to me from some *other* computer-screen. (In such a case, I'd be seeing that other screen, though I might *think* that I was seeing the screen before me.) Nor am I seeing the screen if those experiences are hallucinations.

In this respect, sense-perceptions are like money. For a piece of a paper to be a dollar bill, it has to have a certain provenance. If it was created in somebody's basement, it's a counterfeit bill. Similarly, a given experience *isn't* a perception of X *unless* it has a certain provenance, even if, setting aside its provenance, it is indistinguishable from such a perception.

Granting that I can have an experience *just like* that of seeing a computer screen even if I'm not seeing a computer screen, could I have an experience *just like* that of experiencing intense pain if I weren't experiencing intense pain? No. An experience just like that of intense pain *is* an intense pain. The same thing *mutatis mutandis* is true of instances of any sensation of any kind. An experience *like* that of intense pleasure *is* one of intense pleasure. It's irrelevant how that experience was caused. Where sensations are concerned, the appearance *is* the reality. Anything *like* a sensation of a given kind *is* such an experience.

Sensations aren't the only mental entities that have this property. Anything that, setting aside its mode of origination, is just *like* a belief that 2+2=4 *is* a belief that 2+2=4. To be sure, that belief is a psychological structure; it isn't a constituent of consciousness. And, although one can obviously be conscious *of* one's belief that 2+2=4, one's consciousness of it cannot be identified with, and is to no degree constitutive of, that belief itself. So how it appears to one is distinct from how it is, and it thus cannot be said that, where that belief is concerned, "the appearance is the reality." But what matters in this context is that its being a *bona fide*belief that 2+2=4 has nothing to do with it how it came about.

Whether that belief constitutes *knowledge* that 2+2=4 *does* so depend. It isn't knowledge if it has if the reasoning that led to it is spurious and, therefore, unless it has a certain provenance. So, just as our sensory experiences must have certain origins if they are to be *bona fide*perceptions, so beliefs must have certain origins if they are to constitute knowledge.

This brings us to an important point. Words like "fake" and "real" are relative (in the sense that they denote relations, *not* in the sense that it is to any degree a "matter of opinion" whether a given usage of such a term is legitimate or not). A given piece of paper may not be a real dollar bill. But it's

real in other respects. So it's fake *relative* to the standards that a thing must meet to be a dollar bill. But it's real *relative* to the standards that a thing must meet to have a certain molecular structure, shape, mass, and so on.

A slight detour help clarify this points. To be a mother is for there to be a certain object x such that one bears a certain relation to x. But the word "mother" is *syntactically* very different from expressions, such as "to the north of" or "three feet from", that clearly denote relations, and *syntactically* it's exactly like expressions, such as "a sad person", that denote properties, as opposed to relations. So, whereas "New York is to the north of" is syntactically ill-formed and therefore isn't a sentence, "Mary is a mother" is a perfectly good sentence, and so is "Mary is a sad person." But *logically* "mother" is comparable to "to the north of," not "a sad person." For, whereas one can be a sad person all by oneself, so to speak, one can no more be a mother all by oneself than a place can to the north of all by oneself. Similar points hold of words such as "real", "bogus", and the like. Syntactically, those words are in the same syntactic category as expressions that denote properties. At the same time, they're in same logical category as words that denote relations, the reason being that "X is real" is, if meaningful, an abbreviated way of saying that X bears a certain relation to a certain set of standards.

Thus, a conscious experience that is fake in one respect may be real in another. Even if all of the experiences that we believe to be perceptions are hallucinations and are therefore "fake" perceptions, as opposed to "real" ones, they are still real in other respects. Perceptions give us messages. They tell you that such and such is the case in certain places. Those messages happen to be correct. (In this context, I am using the word "perception" in the way that some authors use the term "accurate perception.") Hallucinations, and other pseudo-perceptions, *also* give you messages. Those messages are either false or coincidentally correct. A false statement is just as much of a statement as a correct one. So, considered as a messages, non-veridical hallucination and coincidentally veridical hallucinations are no less real than their (non-accidentally) veridical counterpart. So, even though, when considered as a *perception*, a given experience's reality doesn't merge with its appearance, its reality *does* so merge when it's considered as a message.

To sum up, where some psychological conditions are concerned, the appearance and the reality merge. By contrast, where knowledge of the *external world* is concerned, the appearance and the reality are always

distinct and are often (in fact, usually) separated by a large tract of space. Even where touch is concerned, the object touched is remote from the events in your central nervous system that mediate your *experience* of the object. So knowledge of at least some facts about one's psychological condition is considerably more direct than knowledge of facts about the external world.

8.2 Some limits to self-knowledge

Notwithstanding what was just said, there are reasons to believe that, in many cases, self-knowledge is as difficult to obtain as knowledge of the external world—that, in fact, knowledge of some facts about our own minds is as theoretical and indirect as our knowledge of sub-atomic events.

Before we can discuss these reasons, we must define a term and we must also make a distinction. The term to be defined is "consciousness." In everyday speech, that word is ambiguous. Sometimes, it simply means "knowledge" or "awareness." For example, in the sentence, "Smith's consciousness of the complexities involved in running a municipality is extremely limited," the word "consciousness" seems to have more or less the same meaning as the term "knowledge" or "awareness."

In philosophical contexts, the word "consciousness" has a different and narrower meaning. In such contexts, it refers to one's immediate experience. (Henceforth, we will use the term "consciousness" *only* in its narrow, philosophical sense.)

This must be understood aright. The term "consciousness" does *not* refer to the *objects* of your experience. Right now, let us suppose, you are feeling a cool breeze. The cool breeze is something that you are conscious *of*. In other words, it is an *object* of consciousness. But that cool breeze isn't identical with your consciousness or with any part of your consciousness. What *is* a part of your consciousness is the sensation that the cool breeze causes you to have. What is also a part of your consciousness is your visual perception of the piece of paper that the cool breeze is blowing across your front lawn. But, to echo what we said a moment ago, the piece of paper itself is not identical with your consciousness or even with a part of it, even though it is something that you are conscious *of* (i.e., it is an *object* of your consciousness).

Here we must make a delicate point. Many constituents of our minds exist even when we are sleeping and even when, more generally, they are not occupying our consciousness in any way. You believe that 1 + 1 = 2. But it is only on very rare occasions that this belief of yours is occupying your consciousness. Nonetheless, you obviously don't *lose* that belief when it's not occupying your consciousness. You still have that belief. This is made clear by the fact that it can readily make an appearance in your consciousness. For example, if I tell you that I'll give you an A in the class if you answer the question "what is 1 + 1?", you will without any difficulty be

able to draw on your belief that 1 + 1 = 2, even though, in all likelihood, that belief was remote from your consciousness prior to my asking you that question.

So that belief, existing as it does when it is outside of consciousness, is not a *part* of consciousness. It is not, as we will henceforth put it, a *constituent* of your consciousness.

Thus, there are instances of knowledge that are not consciousness. And there are instances of consciousness that are not knowledge. The euphoria induced by a drug is an instance of consciousness; it is a component of somebody's immediate experience. But it isn't an instance of knowledge. (This is consistent with the fact that somebody experiencing such euphoria could, and probably would, have knowledge *of* that experience, just as you have knowledge *of* the piece of paper blowing across your lawn).

8.2.1 An important fact about consciousness

One feature of consciousness (as philosophers use that term) is that, if something has consciousness then there is “something it is like” to be that thing. If you are in pain, then there is “something it is like” to be you. So the presence of consciousness in something is sufficient for there to be “something it is like” to be that thing. Further, the presence of consciousness in something is *necessary* for there to be “something it is like” to be that thing.

There is nothing it is like to be a pinecone, since pinecones are not conscious. But if, by some miracle, a pinecone were to become conscious, then there *would* be something it was like to be that pinecone. The presence of consciousness in something is thus both *necessary* and *sufficient* for there to be something it is like to be that thing. In fact, consciousness may be *identified* with a thing’s having the property of there being something it is like to be that thing.

8.2.2 An important fact about your belief that 1 + 1 = 2

These points help clarify our earlier point that your belief that 1 + 1 = 2 is not a *constituent* of your consciousness (even though it may occupy your consciousness, just as the spectacle of litter blowing across a street may occupy your consciousness without, of course, being a veritable *ingredient* of that consciousness). Although there is something it is like to have a pain or a tickle, or to experience a sense-perception, there isn't anything it is like to believe that 1 + 1 = 2.

Of course, that belief may have *effects* on your consciousness—in other words, it may generate constituents of consciousness (without itself being such a constituent)—and there is something it is like to experience those constituents. For example, your having that belief may lead to your getting an A on a test, which in turn may fill you with joy.

8.2.3 A consequence of this important fact: the existence of the unconscious

Thus, your belief that 1 + 1 = 2 is not a *constituent* of your consciousness. It is not like a pain or a tickle or a perception. But it is clearly "conscious" in some sense of the word. (It is not buried in your subconscious; it isn't like the previously discussed processes by which your mind converts disturbances of your sensory surfaces into perceptions.) It is not a *part* of your consciousness. So in *that* sense it is *un*conscious. So to the extent that it is "conscious," it is in the sense that you are conscious *of* it.

What we just said about that particular belief is true of many, if not all, beliefs and, indeed, of many desires, aspirations, regrets, intentions, and emotions. Consider your intention to receive a degree from an institution of higher learning. That intention obviously has effects on what you do. (You choose to show up to class, take exams, and so on.) And it therefore obviously has effects on what is occupying your consciousness. (Because of that intention, you are now reading a book and are therefore having various perceptions and feelings that are constituents of your consciousness.) But that intention cannot be identified with any *constituent* of your consciousness. (That intention continues to exist even when you are asleep.) That intention is better thought of as a structural fact about your mind that *underlies* and *regulates* the events that populate your consciousness, without itself being such an event. Similar remarks hold with respect to (for example) your affection for certain people, your dislike of other people, your confidence in your ability to (for example) play tennis, your doubts about your ability to (for example) play the harmonica, and so on.

These points give us insight into an issue of great importance. You don't *have* to be conscious of anything that you *are* in fact conscious of. You are now conscious of a piece of paper (or computer screen) in front of you. But you don't have to be, and in a few minutes you won't be. You may, at a certain moment, be conscious *of* the fact that you believe that 1 + 1 = 2. But you obviously don't have to be conscious of the fact. In fact, most of the time, you are not thinking about that fact and therefore aren't conscious of it (though you readily become so). So the truth is that we don't have to be, and in fact are usually not, conscious of much of what is in our own minds. This shows that, contrary to what a number of authors have alleged,[250] there is at least some reason to accept the thesis, urged by thinkers in the psychoanalytic

tradition, that we are not conscious of much of what is in our own minds.

Of course, thinkers in that tradition hold not only that we *aren't* conscious of much of what occurs in our own minds, but also that emotional blocks may lead to our having very wrong and distorted and impoverished beliefs about our own minds. Whether that thesis is correct lies outside the scope of this book. But what seems reasonably clear is that there are good grounds for at least *one* contention of that school of thought—namely, that we are, at any given time, unaware of many of the events that constitute our very own mental lives. And, once this is granted, it becomes a garden-variety empirical question whether, unless we take special measures to learn about ourselves, we have more than a superficial understanding of our own minds.[251]

9.0 Knowledge by acquaintance vs. knowledge by description[252]

At the beginning of this chapter, we distinguished between two kinds of knowledge. The one consists of awareness of people (or, more generally, objects); the other consists of awareness of truths. There is reason to think that up to a point, but probably not entirely, the first kind of knowledge collapses into the second. The reason for this is that, with certain very specific exceptions, awareness of objects is identical with knowledge of truths of a certain kind—that, more specifically, it is identical with knowledge of truths to the effect that certain *descriptions* are in fact satisfied.

A story will help clarify what it means to say that a description is "satisfied." In order to collect insurance money, an unscrupulous store owner falsely claims that he is robbed. He gives a description of the (nonexistent) perpetrator of the (non-existent) crime. He goes to the police station and tells the police: "the person who robbed me was a tall male; he had a tattoo that said "I love mom" on his forehead; he had pointy gold teeth . . ." Here the store owner is really saying that there exists an object (a person) that has certain characteristics. To put it another way, he is saying that a certain description is satisfied, that description being "an entity that is a human being who robbed me and has pointy gold teeth and is male and . . ." Any statement to the effect that a description is satisfied is known as an *existence-claim.* So the store owner is affirming the following existence-claim:

(1) There exists an object that is a *person x such that x robbed me, x is a male, x has pointy gold teeth, x has a tattoo*

that says "I love mom" on his forehead . . .

That existence-claim is false, of course, the reason being that the relevant description (the italicized part of 1) is not satisfied. But, of course, many existence-claims *are* true, the reason being that many descriptions *are* satisfied. For example, the following existence-claim is true:

(2) There exists a *person who is currently (in Feb. 2008) the U.S. President.*

And the reason that 2 is true is, of course, that the relevant description (the italicized part of 2) *is* satisfied. (It is satisfied by George W. Bush.)

9.1 Theoretical knowledge as descriptive knowledge

Whenever knowledge is theoretical or otherwise indirect, it is descriptive (i.e., it consists of knowledge of the truth of an existence-claim). We are aware of the existence of electrons. But we don't see them or otherwise sense-perceive them; and we certainly don't have the kind of direct experience of them that we have of some of our own psychological states (e.g., our joys and sorrows). So in what sense are we aware of electrons? The answer: we know the truth of various existence-claims that they satisfy, for example:

> (e) there exist various macroscopic, and therefore perceptible, phenomena that, given known physical laws, are most plausibly explained in terms of theories positing the existence of negatively charged particles orbiting around the nuclei of atoms . . .

We are aware of electrons in the sense that we know the truth of existence-claims that they satisfy. Since, as we discussed, an existence-claim is a statement to the effect that a description is satisfied, our awareness of electrons is entirely descriptive in nature.

Remarks similar to those just made about our knowledge of electrons hold with respect to our knowledge of anything theoretical or, indeed, anything whose existence is inferred, as opposed to known directly.

9.2 Perceptual knowledge as descriptive knowledge

Our knowledge of *directly perceived* objects is strictly descriptive. Consider your current visual perception. What it tells you is that a certain existence-claim is true, that existence-claim being along the following lines:

> (sp) There exists a piece of paper that is in such and such a location with respect to me, that has such and such a shape and color, that has such and such discolorations on it . . .

Perception is description. It is inconceivable that one should have a perception that didn't describe the world in a certain way and that did not, therefore, say of the world that it contained an object having certain characteristics. (Even if you are looking into empty space, your perception is to the effect that the description *stretch of empty space* is satisfied by at least one entity in the world.) Thus, a given perception tells you that some existence-claim is satisfied, and perceptual awareness is therefore descriptive in nature.

This shows that awareness of sense-perceived objects, and therefore of other people, is identical with knowledge of truths of a certain kind (existence-claims), vindicating our earlier suggestion that knowledge of objects at least sometimes collapses into knowledge of truths.

Are there any cases where knowledge is *not* descriptive? The answer seems to be "yes." If you have an intense pain, it seems that your awareness of that pain is utterly direct and doesn't involve your knowing that some existence-claim is satisfied. The pain is right there, so to speak, in the theatre of your consciousness, and it needn't be represented to you through the intermediary of some description. So, at least arguably, we have non-descriptive knowledge of at least some of our psychological states. But, for the most part, it seems that our knowledge is descriptive.

Why, in the penultimate sentence, do I say "at least arguably"? Because *having* pain doesn't necessarily involving *knowing* that one is pain. Knowing that one is in pain involves more cognitive wherewithal than being in pain. And it may be that the steps one must take to go from being something that merely experiences pain to being something that *knows* it's in pain inevitably makes one's knowledge of one's pains indirect. I myself think that, all of this notwithstanding, one *does* have direct knowledge of one's own conscious,

phenomenal states; for I think that, unless we had direct knowledge of such states, we couldn’t have any spatiotemporal knowledge. Nonetheless, given the distinction between experiencing pain and knowing that one is experiencing pain (or any other phenomenally pregnant, conscious condition), it is not a trivial question whether, indeed, one’s knowledge of the fact that one is in pain (same qualification) is direct.

9.3 Procedural vs. declarative knowledge

We have found that, at least arguably, there are *some* cases of awareness that are not awareness of truths. At least arguably, we can be *directly* aware of some of our own psychological states and can therefore be aware of them without being aware of truths that describe or otherwise implicate them. Are there other kinds of awareness or of knowledge that are *not* awarenesses of truth?

Some philosophers and psychologists say "yes," the reasoning being as follows:

> (KH)[253] We must distinguish knowledge-that from knowledge-how. I know *that* Paris is the capital of France. This is knowledge of a truth. It is knowledge-*that.* I know *how* to ride a bike. This is knowledge, but it isn't knowledge of a truth. It is knowledge-*how,* not knowledge-that.
>
> Any *skill* is a case of knowledge-how. It's easily shown that knowledge-how and knowledge-that are fundamentally different things. No matter how much I know about the physics and geometry of bike-riding, I may not be able to ride a bike at all; and no matter how good I am at riding a bike, I may not be able to tell you the first thing about the geometry or physics involved.
>
> No amount of knowledge-that adds up to a skill, and no amount of skill adds up to knowledge-that. So knowledge-how is not knowledge-that, but it is, of course, knowledge of some kind or other.

But it's a spurious argument for a false conclusion. Knowledge of a language is knowledge-how. To be able to speak English is to have a skill. But that skill involves having knowledge-that, for you must know various syntactic and semantic truths (e.g., the truth that "snow" picks out snow) to speak English. Knowledge of a language is operationalized semantic and syntactic knowledge. In general, procedural knowledge is operationalized declarative knowledge. (By "declarative knowledge," I mean knowledge of truths.)

Of course, not all procedural knowledge starts out as declarative

knowledge. (Your ability to ride a bike didn't start out as non-procedural knowledge. It was procedural all along.) Moreover, one has no conscious access to the declarative knowledge internal to one's procedural knowledge. (One has conscious access only to analogues of that knowledge. A given piece of knowledge may be stored at many different cognitive levels. Knowledge is oftentimes "over-represented." A baseball player who is also a physicist may have two distinct pieces of knowledge of some one fact.) Contrary to what Ludwig Wittgenstein (1889–1951) and John McDowell (1942–) hold, the conclusion to be drawn from these facts is *not* that knowledge-that is no part of knowledge-how, and is instead that the mind is compartmentalized, and the right hand doesn't always know what the left is doing.

10.0 Intuitive vs. discursive knowledge

Whenever we know something without knowing *how* we know it, we have *intuitive* knowledge. Whenever we know something and we *do* know how we know, we have *discursive* knowledge. So to know something intuitively is to "just know it," whereas to know something discursively is to know it *while also* knowing the grounds on the basis of which one knows it.

Some stories will clarify these points.

> Story #1. Timmy, who is 10 years old, is in a classroom and he is looking at a figure on the blackboard. The teacher points to the figure and asks Timmy "what shape does that have?" Timmy correctly says "it's a circle." The teacher asks: "how do you know? Why is it a circle as opposed to oval?" Timmy says: "I just know. It's just obvious that it's a circle."

> Story #2. Jones, who is a math major, is in a classroom and he is looking at a figure on the blackboard. The professor points to the figure and asks Jones "what shape does that have?" Jones says "it's a circle." The professor asks: "how do you know? Why is it a circle as opposed to oval?" Jones says: "because, if *x* and *y* are any two arbitrarily short segments of that shape's periphery, *x*'s degree of curvature equals *y*'s."

Commentary: Obviously both Jones and Timmy know that the figure on the board is a circle. But Jones knows something that Timmy does not know. But what, exactly? A possible answer is this: "Timmy doesn't know why the figure is a circle, whereas Jones does have that knowledge." But surely Timmy's belief is based on the very fact about the figure's shape that Jones has described. It seems to me that—in some way, at some level—Timmy is aware of the fact about that figure that Jones is describing. He doesn't know it in the same way as Jones—but he still knows it. So it is wrong, or at least very misleading, to say that Timmy doesn't know what it is about that figure that makes it a circle.

10.1 How intuitive and discursive knowledge differ

Here's what I think. Jones does, whereas Timmy does not, know *on what grounds he himself is judging the figure to be a circle.* Timmy and Jones both believe that the drawing is a circle, and (so we may assume) each believes it for the right reasons. But Jones knows, whereas Timmy does not know, what the considerations are that are moving him to make that judgment. The difference between Timmy and Jones is that Jones has a kind of self-knowledge that Timmy lacks.

10.2 How intuitive and discursive knowledge differ (continued)

Here's another hypothesis (which, I think, is compatible with the one just given). Even though both Jones and Timmy know that the drawing is a circle, Jones knows this in *two distinct ways,* whereas Timmy knows it in only one way. Timmy has a strictly visual (or, in any case, sensory) knowledge of this fact, whereas Jones has what might be described as *propositional* knowledge of this same fact. Let us now develop this point.

First of all, a proposition is what is meant by a sentence—it is the information encoded in a sentence. The information provided by sentences has two striking features: first, it is very schematic; second, it is the kind of information that can be grasped only through an act of *judgment,* as opposed to perception. (These two facts are related, in ways that will now become apparent.)

Suppose you see a certain man, whom you know to be Smith, next to a certain telephone pole, and you report that fact by saying:

(SP) "Smith is next to that telephone pole."

Compare what you actually *see* to SP. You don't just see *Smith.* You see an individual with very specific characteristics (a specific shape, a specific posture or series of specific postures, wearing specific clothes, having hair of a specific color, and so on). And you don't just see a *telephone pole.* You see an object having a specific shape, in a specific location, with a specific color (and with specific discolorations on it), with wires of a certain kind attached to certain parts of it, and so on. Thus, the sentence provides a very schematic and abstract representation of the information provided to you by your visual perception. So, from one viewpoint, the sentence *under*-represents the content of your visual perception.

But, from another viewpoint, the sentence *over*-represents the content of your perception. Given only what you see—given only what your eyes are telling you—it could be that you are seeing somebody other than Smith; it could be that you are seeing somebody who looks just like him but isn't him—or even that you aren't seeing a person at all and are in fact seeing a robot. When you characterize that thing as "Smith," you are making a judgment as to who and what that thing is; you are making a judgment as to the biological species and as to the specific identity of that individual. Though *partially*

based on what your eyes are now telling you, that judgment incorporates information that is not derived from that specific perception—information that is derived from past meetings with various individuals and from inferences drawn therefrom.

And given only what your eyes are telling you, you don't know that you are seeing a *telephone* pole. A telephone pole is something that serves a specific function within a complicated network of social and physical relations. When you are staring at the telephone pole, you cannot literally *see* it performing these functions. Thus, implicit in your characterization of that thing as a "telephone" pole is a judgment as to its social function. And, though partially derived from your sense-perception, the grounds for that judgment have roots in past experiences of yours and also in inferences drawn therefrom. So, from that viewpoint, SP over-represents what your eyes are telling you on this occasion.

Similar points hold with respect to the differences between Timmy and Jones. Jones isn't simply seeing a certain object's shape; he isn't simply uploading perceptual information. He is making a judgment about that information. (So Jones is generating "meta-information.") By contrast, Timmy is just uploading the information.

A related point is that, although both Timmy and Jones know that the drawing is a shape, Jones knows this fact in more ways than Timmy. This can be understood in terms of our earlier point that knowledge is descriptive. Timmy knows that the drawing is a circle and so does Jones. But, in knowing of the existence of this one *fact*, Jones knows two different *truths*. He knows the truth embodied in some graphic (visual) description of that fact, and he also knows the truth embodied in some non-graphic representation of that fact. By contrast, Timmy knows the first truth but not the second.

We thus seem to have two answers to the question "what does Jones know that Timmy doesn't?" First, Jones has a certain kind of self-knowledge that Timmy lacks: Jones does, whereas Timmy does not, know what the considerations are that led him to believe (rightly) that the drawing is a circle. Second, Jones knows the fact in question under two different descriptions (one pictorial, the other conceptual), whereas Timmy knows it only under one description.

10.3 The paradox of analysis

These points may give us some insight into an old philosophical dispute. Consider the statement:

> (CR) *x* is a circle if and only if *x* is a two-dimensional figure of uniform curvature.

Obviously CR is informative. It isn't like the statement:

> (CC) *x* is a circle if, and only if, *x* is a circle.

CR is an example of a *conceptual analysis.* A conceptual analysis is a statement that lays bare the structure of a concept.

According to G.E. Moore, it is paradoxical to suppose that conceptual analyses can be informative. There is, he says, a "paradox of analysis." Moore's reasoning seems to be this. The sentence "water is H_2O" expresses an analysis of water. In that sentence, the words flanking the "is" refer to the same substance. If they didn't, that sentence wouldn't express a correct analysis. For the same reason, the expressions flanking the "iff" in a sentence expressing a conceptual analysis must refer to the same concept. But if they did, then CR would be trivial, like CC, and thus fail to be an analysis.

This is not good reasoning. True—"water is H_2O" expresses a "chemical analysis"; and, in it, the two words flanking the "is" co-refer. True—CR expresses a "conceptual analysis." But it doesn't follow that, in it, the expressions flanking the "iff" must co-refer. Moore's reasoning is comparable to that of somebody who thinks that, since so and so isn't interested in having "legal" (law-related) discussions, so and so has no wish to have "legal" (law-permitted) discussions.

In any case, Moore's paradox is easily enough dissolved. The truth that one grasps in virtue of knowing that:

> (i) *x* is a circle,

is different from the truth that one grasps in virtue of knowing that:

(ii) *x* is a two-dimensional figure of uniform curvature.

Both of those truths describe the same *fact*—the fact that a certain object has a certain shape. But a given fact can obviously be described in different, but equally correct ways. The fact that *x* is a triangle can be described by saying that *x* is the area bounded by three straight edges, any two of which intersect, but not all of three of which do so. And that same fact can be described by saying that *x* is a three-sided, straight-edged, closed, planar figure. And:

> (iii) *x* is a triangle iff *x* is the area bounded by three straight edges, any two of which intersect, but not all three of which do so iff *x* is a three-sided, straight-edged, closed, planar figure,

is just another way of saying that there are three different ways of describing *x*'s shape.

Facts aren't truths. One can't know this without knowing how to solve the Paradox of Analysis. But Moore didn't know it.

11.0 Justification vs. discovery

The question:

> (1) What events in Newton's mind led to his coming up with his theory of gravitation?

must be distinguished from the question:

> (2) What merit does that theory have?

One can know the answer to the one question without knowing the answer to the other. To know why Newton had such and such beliefs, I don't need to know whether those beliefs are correct or otherwise meritorious, and to know whether those beliefs are correct or otherwise meritorious, I don't need to know why Newton had them.

Given a belief, the question "is it correct?" is to be distinguished from the question "how was it arrived at?" and knowing the answer to the one question doesn't necessarily involve knowing the answer to the other.

"Justification isn't discovery." Philosophers of science often say this. In so doing, they're affirming the points just made.[254]

11.1 The connection between justification and discovery

Granting that an idea's pedigree is to be distinguished from its merits, it's misleading to say that justification isn't discovery. To discover something is to come to know it. Knowing something involves having a justified belief in it. The justification for that belief will be encoded in the thoughts that led to that belief. No justification, no knowledge. No knowledge, no discovery. Thus, justification *is* discovery—or at least a good part of it.

We must distinguish *discovering* something from merely *coming to believe it.* One cannot discover something without coming to believe it, but one can come to believe something without discovering it. Coming to believe something doesn't necessarily involve being justified in believing it or, therefore, in knowing it or, therefore, in discovering it.

But to the extent that a person's thought-processes are expressions of intellectual ability, and not of prejudice or some other discovery-thwarting factor, that person's beliefs are likely to be justified and, consequently, that person won't come to believe something without being justified in doing so. Given that the conjectures that led to great scientific and intellectual advances are obviously reflections of intellectual ability, as opposed to blind luck, it follows that oftentimes the justification for such a conjecture *is* to be found in its psychological antecedents. It follows that justification is inherent, not just in scientific discovery, but also in scientific *proto*-discovery. In other words, it's inherent not only in the processes that lead to advances, but in the processes that lead to the conjectures that lead to such advances.

Great ideas often consist of many intricately interconnected subsidiary ideas. The odds of accidentally stumbling upon those ideas, as opposed to arriving at them on the basis of knowledge and insight, are exceedingly slim. Unless Newton and Keynes were just lucky, which is unlikely, the thoughts they had along the way to making their discoveries were driven by an understanding of how the evidence at their disposal ought to be interpreted. Those thoughts were therefore to the effect that, *since such and such evidence justifies thus and such contentions, thus and such contentions ought to be accepted; therefore, thus and such contentions are true.* So, setting aside cases where the discoverer was simply lucky, the psychology and the logic of discovery mirror each other.

Pascal said "it is through intuition that we discover; it is through logic that we prove."[255] This is wrong. It is through logic that we discover and it is

through logic that we prove. For there is no discovery without justification. An unjustified intuition isn't a discovery at all. And if an unjustified intuition happens to be right, it's a matter of luck. While some non-trivial theoretical innovations may be made through luck, most of them are not. (I doubt that any of them are.) And a correct belief arrived at through intuition, as opposed to luck, is *ipso facto* justified. Proof is the identification of the logic embodied in discovery.

11.2 Discovery vs. prediscovery

“For all we know,” many a philosopher of science says,

> (BL)[256] It was Newton’s being hit on the head by an apple that led him to posit the inverse-square law. So, were we to know exactly what psychological processes led to Newton's putting forth this conjecture, we wouldn't necessarily find a justification for it; and should we find such a justification, our doing so would be a matter of blind luck. There's no point in poking around Newton's psyche unless you're interested in his relationship to his mother. If you're interested in physics, you'll find nothing.
>
> If you psychoanalyze great scientists, you'll find *causes*, not *justifications*. You'll find the *cause* of Einstein's belief that inertial and gravitational mass are the same thing. But you're no more likely to find a *justification* for that belief than you are if you look in Einstein's sock-drawer. Psychology isn't logic. Justification isn't discovery.

BL is overstated. Not everybody who is hit on the head becomes an innovator in physics. So far as a blow to the head had anything to do with Newton’s discoveries, it was only by way of its triggering an intricate train of thought. We must distinguish the instigators of a thought process from the thought process itself. Being hit on the head was no part of Newton’s discovery process. It was, at most, a part of his discovery *pre*-process. The blow, supposing that it happened, merely activated the relevant processes and therefore wasn’t constitutive of them.[257]

11.3 The roots of the problem

The connection between discovery and justification is easy to overlook if one doesn't bear in mind the following four points:

> (1) Having a justification for a belief doesn't always involve knowing what that justification is.
>
> (2) Knowing that justification doesn't always involve being able to say what it is. (One can't always articulate what one knows.)
>
> (3) Being able to articulate that justification is different from actually doing so and, in so doing, transmitting that justification to others.
>
> (4) Transmitting *one's own* justification for a given belief is different from producing a *general* justification for it. (It is one thing for *me* to have certain experiences that give me an air-tight justification for believing *P*. But given that those experiences may have occurred under singular circumstances, the likes of which are unlikely to be repeated, it is quite another thing for me to provide others with a way in which they can become justified in believing *P*.)

11.3.1 (1) explained

There is often a delay between having a justification for a belief and knowing what that justification is. My intuition tells me that so and so is ill-intentioned, even though, on the surface, he is behaving amicably; that a given sentence is grammatical, even though my grammar book classifies it as ungrammatical; that a given animal is about to attack, even though it isn't giving off any of the obvious signs.

We trust our intuitions, and our confidence in them is usually vindicated. (Speaking personally, I always regret second-guessing my intuitions.) But when we have them, we don't know *why* they are to be trusted. If we did, they wouldn't be intuitions, since an intuition is, by definition, a belief whose justification, if it exists, is unknown to the person having the belief. Some have said that, for this very reason, intuitions are never knowledge. "If one has grounds for believing something," it is said, "one knows what those grounds are."

This reasoning couldn't possibly be more off the mark. It is only to the extent that one hasn't fully internalized the relevant principles that one's thought *fails* to be intuitive. It is only pedants and neophytes who rely on discursive thought. Good pianists don't have to *think* before they hit every note. Bad pianists do. Good pianists, unlike their bad counterparts, have internalized the relevant rules and, for that reason, don't need the crutch of conscious, discursive mentation. Relying on consciously executed operations to apply rules is like relying on an instruction manual.

Discursive thought is remedial thought, and thought is non-remedial only insofar as it's intuitive. It doesn't follow that all non-discursive thought is truth-conducive. But it does follow that not all intuitive thought is bad. Given that knowledge is intuitive only in so far as its possessor doesn't know the justification for it, it follows that one can be justified in having a certain belief without knowing what that justification is.

Nonetheless, philosophers of science tend to hold that one has a justification only when one states, or at least can readily state, one's justification.[258] For this reason, they're forced to say intuitions are never justified. Given that scientific breakthroughs result largely from intuitive thought and often themselves take the form of intuitions, philosophers of science have no choice but to see the generation of discoveries as having no connection, or only a fortuitous connection, to the eventual justification of

them.

11.3.2 (2) explained

One can know one's reason for having a given belief without being able to state it. A skillful interrogator may know why he believes that the suspect is lying but not be able to say why he thinks this. His judgments are made on the basis of subtle cues that he might have a hard time articulating. Articulating one's knowledge may involve translating it from one medium into another. This is the case with the interrogator. His judgments are responses to visual information. Articulating those judgments would involve translating that information into a non-visual medium.

Imagine the following. Max believes that metal usually expands when heated, and this belief of his is thoroughly justified. (He's done numerous experiments, absorbed copious amounts of theory, and so on.) But for some reason Max refuses to *state* his reasons for having that belief. Max's situation is very different from the interrogator's. Unlike the interrogator, Max knows what the justification for his belief is. In any case, even though Max's belief is thoroughly justified, it could be said that Max has "failed to justify" it. Thus, the expression "unjustified belief" is ambiguous. On the one hand, it can refer to beliefs that aren't justified. On the other, it can refer to beliefs the justifications for which, if they exist, haven't been stated.

11.3.3 (3) explained

Even though a given scientist has cogent reasons for accepting a given hypothesis, and makes every effort to communicate them to others, it doesn't follow that others know what those reasons are. Others may not understand what that scientist is saying—maybe his theories are too advanced for them. Maybe they do understand what he's saying but haven't yet had a chance to authenticate his results. Until they have good reason to accept the scientist's testimony, *his* justification for accepting the theory in question isn't yet theirs.

When epistemologists say that a belief is justified, they mean that *the person who has it* is justified in having it. When philosophers of science say that a belief is justified, they mean *either* what epistemologists mean *or*, what is more likely, they mean that people *in general* are justified in having it.

There was a time when many people believed that smoking causes lung cancer, but when only a few researchers actually knew it, since those researchers alone knew the relevant clinical data. During that time, it would have been correct to describe "*the* widespread belief that smoking causes cancer" as unjustified.

In general, there are many cases where *the* belief in such and such is unjustified and therefore isn't knowledge, even though specific scientists had already discovered such and such and therefore knew it, and were therefore justified in believing it. It thus happens that such and such is discovered before *the* belief in it is.

Chapter 11 Skepticism and the Justification of Inductive Inference[H7]

1.0 What is skepticism?

A skeptic is a doubter, not a disbeliever. For me to be skeptical about Smith's assertion that he's an astronaut is for me to doubt that he's an astronaut. It isn't for me to *dis*believe that assertion—only to question it. If I'm convinced that Smith is not an astronaut, I'm not *skeptical* about his claim. Skepticism is doubt, not disbelief.

But there is some tendency for skepticism to give way to disbelief. Even if I never definitively learn that Smith is *not* an astronaut, if I continue to find no good reason to believe that he is one, I'm more likely than not to disbelieve it. We feel that, if a claim is true, some positive evidence for it will turn up: the mere absence of counterevidence isn't enough for us.

Does our human tendency to transition from doubt to disbelief have any logical basis? Or is it just another case of emotion and instinct corrupting rationality? It depends on the case. In some cases, but not all, failure to find positive evidence is itself counterevidence; and, in such cases, our tendency to slide from doubt to disbelief does have a logical basis. But, as we'll see in this very chapter, unless some highly defensible forms of skepticism can be refuted, this claim of mine is simply wrong.

Many important debates in philosophy concern the merits of some form of skepticism. Are there non-spatiotemporal entities? (Skepticism about platonic entities.) Do we have beliefs, feelings, etc., that we aren't aware of having? (Skepticism about unconscious ideation.) Could observation ever provide adequate grounds for believing in unobservables (e.g., electrons) or should we regard terms like "electron" as referring to fictions of some kind or as not referring to anything at all? (Skepticism about theoretical entities.)

As these examples indicate, philosophical skepticisms involve whole *categories* of truths. One doubts *all* truths affirming, or presupposing, the existence of non-spatiotemporal entities, unconscious emotions, theoretical entities, etc.

In this chapter, we'll discuss what are probably the two most important forms of skepticism: *skepticism about the external world* and *skepticism about inductive inference*.

To be a skeptic about the external world is to doubt that what our senses are telling us is accurate. One is a skeptic of this kind if one believes there to be no good reason to believe that one's sense-perceptions are anything other than hallucinations.

A bit of background is needed to say what it is for one to be a skeptic about inductive inference. Some knowledge is direct. One doesn't have to figure it out—one just knows it. Somebody who is experiencing a searing pain knows it, and he didn't have to figure it out. But much knowledge is indirect. Einstein didn't "just know" that a body's mass increases as its speed increases; he had to figure it out. In other words, he had to *infer* it. So for knowledge to be indirect is for it to be acquired through inference.

There are two parts to any inference: the premises and the conclusion. The conclusion is the belief arrived at—the thing that, when the inference leads from the right premises to the right conclusion in the right way, is "figured out." The premises are the pre-existing beliefs or assumptions on the basis of which one arrives at the conclusion.

There are two kinds of inferences: deductive and inductive. An inference is deductive if, supposing one's premises true and one's reasoning non-faulty, it is impossible for one's conclusion to be false. If my premise is that *x* is a square and my conclusion is that *x* has four sides, it's logically impossible for my conclusion to be false if my premise is true. So this would be a case of a deductive inference. An inference is inductive if, supposing one's premises true and one's reasoning non-faulty, it *is* possible for one's conclusion to be false. If my premise is that, in the 35 years I've been drinking tap water, I have never become sick as a result of doing so, and my conclusion is that I won't become sick as a result of drinking the glass of tap water in front of me, my conclusion *may* be false, even though, so we may assume, that belief was generated in an utterly reasonable manner. So this would be a case of an inductive inference.

A skeptic about inductive inference holds that it is impossible, as a matter of principle, to learn anything through non-deductive inference. He holds, in other words, that one cannot acquire knowledge through inductive inference. All knowledge is either direct or is arrived at through deductive inference.

If this is right, then we know only an infinitesimally small fraction of what we think we know. This is because, for reasons thoroughly discussed in Chapter 10, no knowledge about the external world is direct, and only some knowledge of ourselves is direct.

We will start by discussing skepticism about induction.

2.0 Skepticism about induction

David Hume made it clear why it may be doubted whether inductive inference ever leads to knowledge. Here is Hume's argument[H8] :

> (HA)[259] Consider some inductive inference—for example, "the sun has always risen in the past; therefore, it will rise tomorrow." First of all, this argument is not *deductive*, since, even if the premise is granted, it's a theoretical possibility that the conclusion will be false. The sun might explode, after all. So the conclusion is, at best, probable. That's step 1.
>
> Here's step 2. Given that the connection between premise and conclusion isn't deductive, we must ask: *how* does the premise warrant the conclusion? In other words, what reason does the (supposed) truth of the premise give us for accepting the conclusion?
>
> Here's an answer that seems tempting at first: "The past resembles the future; the known resembles the unknown. Nature is uniform. So given that such and such has been the case until now, we may infer that such and such will continue to be the case. Given that, until now, the sun has risen, pinewood has floated, water has quenched thirst, etc., it's reasonable to conclude that the sun will continue to rise, and so on. It is thus the principle that nature is uniform—or, as we'll call it, the *uniformity principle* (UP)—that allows us to make rational, non-deductive inferences about the unknown on the basis of the known."
>
> But this argument is no good. No one directly knows that UP holds for all places and times. At most, one knows that it *has* held for the fragments of nature with which one has personally been acquainted. One doesn't directly know that it held of past objects with which one was not personally acquainted; and one doesn't directly know that it will hold of future objects with which one won't be personally acquainted.
>
> Thus, any knowledge that one has of UP must be inferential. The inference in question cannot be deductive. Given only that the ravens thus far known to one have been black, it doesn't

follow *deductively* that ravens not (yet) known to one are also black. In general, given only that the phi's known to one are psi's, it doesn't follow deductively that the same is true of the phi's that aren't known to one. If it follows at all, it follows inductively, not deductively. But it follows inductively only on the condition that UP is granted. Since, however, a *precondition* for UP's legitimacy is the legitimacy of inferences of the form *since known phi's are psi's, so are unknown phi's*, it follows that UP is not legitimate and, therefore, that there is no way to justify inductive inference. [260]

3.0 Problems with Hume's argument

Nobody denies that inductive inferences are knowledge-conducive and, therefore, that the conclusion of Hume's argument is false. Even Hume said that, in his bones, he couldn't accept it. But there's no accepted response to Hume's extremely clever argument. Nonetheless, Hume's argument is refutable and a refutation of it must lie in some argument not totally unlike the following.

There are two serious errors in Hume's argument. First, it assumes that the primary form of non-deductive inference is *induction by enumeration.* In actuality, induction by enumeration is less important than, and subordinate to, a very different form of non-deductive inference known as *inference to the best explanation.* (The italicized terms will be defined shortly.) Second, Hume assumes that non-deductive inferences go through only if UP is granted (i.e., only if it's granted that nature is uniform.) But UP is not the relevant principle. The relevant principle, I propose, is that *good explanations minimize causal anomalies*. Let's refer to this principle as "MC" (short for "minimization of causal anomalies"). Another way of stating MC is to say that, given two otherwise comparable theories, the one that does the most in the way of anomaly elimination is the better one.

Unlike UP, MC is, I submit, *analytically* true: it is inherent in the concept of what an explanation is that, other things being equal, the theory that eliminates the greater number of unexplained explainers—and, in particular, the greater number of unexplained causal connections—is the one to be preferred. Because this principle, unlike UP, is analytically true, it doesn't have to be validated on inductive grounds. This makes it possible to validate inductive inference in a way that, unlike the way that *Hume* considers, is not viciously circular.

3.1 Inference to best explanation vs. enumerative induction

There are two forms of inductive inference: induction by enumeration and inference to the best explanation. (See Chapter 12.) An induction by enumeration has the form: every *F* that I've ever known of was also a *G*; therefore all *F*'s are *G*'s. Example: All the balls that I've thus far removed from the urn have been white; therefore, all the balls in (or removed by me) the urn are white.

An inference to the best explanation has the form: "It is known that *S*; it isn't yet known that *S**; but if *S** were true, it would, without doing excessive violence to what we (think we) know about the world, explain why *S* is true; moreover, if *S** weren't true, it would be hard to find a viable alternative explanation of *S*; therefore, *S** is true."

For example: I'm very careful with my credit cards (I never use them to make online purchases, I never let anyone besides reliable merchants see them, etc); and, before my friend Larry moved in with me, no unauthorized purchases were ever made with any card of mine. But after he moved in, my cards were used to make a series of unauthorized purchases. Moreover, the items bought were ones that, for work-related reasons, Larry desperately needs and that nobody outside of Larry's rather unusual profession would have the slightest use for. Knowing all of this, and not knowing any other way to account for the unauthorized purchases, I conclude that Larry made those purchases.

It's widely assumed among philosophers that induction by enumeration is the more fundamental form of inference. This is a mistake. All cases of induction by enumeration, when legitimate, are parasitic on cases of inference to the best explanation. There are, of course, many cases where *one* of my reasons for thinking that all *F*'s are *G*'s is that every *F* I've ever known of was also a *G*. But, setting aside beliefs arrived at through spurious reasoning, there are no cases where that was my *only* reason for believing that all *F*'s are *G*'s. Hume's argument focuses entirely on induction by enumeration. Hume is right, I believe, that *this* form of inference is unjustifiable. But we don't use this form of inference. In any case, we don't use it in the way that Hume thinks.

Moreover, we shouldn't use it and we needn't use it. In some cases we may *think* that we're doing straight induction by enumeration. But, in almost all such cases, a little probing always shows that many interconnected

background assumptions are at work and that, when these are made explicit, what we're really trying to do is to produce an inference to the best explanation. And, although there are cases of straight induction by enumeration, they tend to occur in rather artificial contexts and that, when they do occur, they're *clearly* spurious.

There is a well-known inductive error known as the "gambler's fallacy." (Sometimes it's referred to as the "Monte Carlo" fallacy.) You are playing a game of roulette. (This game involves rolling a ball into a spinning wheel, which is divided into numbered slots.) [H9] You know that nothing about the game is in any way rigged—that ball isn't weighted or otherwise propertied in such a way as to favor a particular outcome, and that neither the wheel nor the underlying apparatus were built or modified to favor any given outcome more than any other. And you thus know that, given only the physical properties of the game, the ball, once rolled, is no more likely to land in this as opposed to that slot. (How you know this is irrelevant; we'll just assume that you do.) You *also* know that, given any slot on the wheel, there is nothing about you that makes a ball you've rolled more likely to land in that spot than in any other—nothing about your body, mind, or ball-rolling practices that favors this as opposed to that outcome. (Again, *how* you know this is irrelevant; we'll just assume that you do.) You've rolled the ball seven times. Each time, at least it has landed in slot #27. It must be stressed that you know there to be nothing about the game or about you that predisposed this to happen—that you know this to be entirely a matter of luck.

Under these circumstances, it would be natural to predict that you'll roll a 27 next time. But that prediction would be irrational: you have no more reason to believe that you'll roll a 27 than you will any other number. And if, by chance, you *do* roll a 27 next time, it will be natural to predict that you'll roll a 27 the *following* time. But this too would be completely irrational. So long as you know that there's nothing inherent in the game or in yourself that favors this as opposed to that outcome, you are never any more entitled to bet that, with the next rolled, you'll roll a 27 than you are to bet that you'll roll any other available number. It doesn't matter how many consecutive 27s you rolled prior to that roll: it could have been ten, a thousand, or a million.

By the same token, no matter how many consecutive 27s you roll, you are never any more entitled to bet that, next time, you *won't* roll a 27 than you are to bet that you won't roll any one of the other available numbers. It would be natural to think otherwise, of course. Given how improbable it is that you

will roll five consecutive 27s, let alone 50, you may have an urge, when rolling for the 51st time, to think:

> *This time,* surely, I'm even *more* unlikely to roll a 27 than I am to roll any other available number. The game isn't rigged. The likelihood of my rolling 51 consecutive 27s is infinitesimal and, therefore, is much smaller than 1/n (where n is the number of available slots). So a 27-roll is *especially* unlikely—more unlikely than a 5-roll or a 13-roll.

But this is spurious reasoning. You are no more, and no less, *un*likely to roll a 27 than you are to roll any other number. With every roll, the chances of rolling any one of the available numbers are identical with the chances of roll any one of the others. Since, by hypothesis, the conditions that obtain on any two occasions are identical, it would be irrational on any given occasion to think any given outcome more likely than any other given outcome.

3.1.1 When is enumerative induction rational?

When *would* it become rational to believe that, next time, you're more likely than not to roll this as opposed to that number—that, for example, you are especially likely to roll a 27? This belief becomes rational when, and *only* when, you have reason to believe that a 27-roll is favored by the structures involved in the game. And that belief, in its turn, is rational if you know that circumstances at all like the following obtain:

> *The ball is magnetically attracted to the 27-slot.
>
> *On any given occasion, you (unconsciously) intend to roll a 27, and you are so talented at rolling marble-sized balls that you can roll a 27 if it is your (unconscious) intention to do so.
>
> *The 27-slot is much bigger than any of the other slots. In fact, it takes up so much space on the roulette wheel that the remaining spaces are too small for the ball[H10] to fit into them.

You are rational to believe that you'll continue to roll 27s to the extent that your having thus far rolled multiple 27s in a row gives you reason to believe there to be some underlying structure favoring that outcome. And to the extent that a long run of 27-rolls *doesn't* give you such a reason, you are *irrational* to believe that you're any more (or any less) likely to roll a 27 than you are any other number. So, no matter how many consecutive 27s you roll, if you know with complete certainty that nothing about either yourself or the set-up of the game favors that outcome, it is irrational to think that, next time, you're any more (or less) likely to roll a 27 than you are a 5 or a 32. Put pedantically, it is only insofar as you have reason to believe in such a structure that you have reason to expect something that has the property of being a ball rolled by you to have the property of landing in the 27-slot.

3.1.2 Generalizing these points

Your knowing of many phi's that are psi's and of none that are not doesn't *necessarily* give you any reason to believe that the next phi you encounter will be a psi; it gives you such a reason *only* insofar as it gives you a reason to believe in some structure or mechanism that disposes phi's to be psi's. If you know on independent grounds that there is no such mechanism, no run of phi's that are psi's, no matter how long, gives you a reason to think that the next phi will be a psi; and you are guilty of the gambler's fallacy—and thus, as we saw, of the most rank absurdity—so far as you think otherwise. Thus, *in and of itself* induction by enumeration is worthless; it's a fallacious form of reasoning and falls in the same category as every other argumentative fallacy. Of course, if you know of many phi's that are psi's and of none that are not, you *do* have good reason to hold that, other things being equal, the next phi you encounter will be a psi. But this is because your having this knowledge gives you reason to believe there to be a principled or lawful connection between a thing's being a phi, on the one hand, and its being a psi, on the other. Here, then, is the *real* structure of inferences that are wrongly thought of as cases of induction by enumeration:

> (i) you know of many phi's that are psi's and of no phi's that are non-psi's;
>
> (ii) on the basis of step (i), you believe there to be some principled connection between a thing's being a phi and its being a psi;
>
> (iii) if you know of a given thing that is a phi, but don't yet know whether or not it's a psi, you know, given (ii), that it's likely to be a psi; therefore,
>
> (iv) you conclude (if only tentatively) that it *is* a psi.

Thus, any case of induction by enumeration that isn't an instance of the gambler's fallacy involves the positing of some mechanism or law that, were it to exist, would explain a certain concomitance—it involves, in other words, a case of *inference to the best explanation.* The best explanation of the fact

that all known phi's are psi's is that, thanks to some mechanism or, in any case, principled connection of *some* kind or other, a thing's being a phi disposes it to be a psi; and as soon as such a mechanism has been posited, but not one second sooner, your knowing of a given thing that it's a phi is a reason for you to hold that it's a psi. So any non-fallacious "induction by enumeration" is nothing other than a case of inference to the best explanation (or, more precisely, is a case of applying some theory that embodies an inference to the best explanation).

Hume's argument assumes that it is only through induction by enumeration that the past is any guide to the future. It assumes that, so far as we have any reason to believe that future phi's will be psi's, it is that past phi's have been psi's. But this assumption is dead wrong. The fact that past phi's were psi's, is not, in and of itself, reason to hold that future phi's will be psi's; it is such a reason *only* to the extent that it suggests some mechanism that disposes phi's to be psi's. Hume's argument assumes otherwise and is thus guilty of assuming the legitimacy of the spurious form of inference involved in cases of the gambler's fallacy. And, if it is to have a chance of going through, Hume's argument must be so reconstructed that it no longer embodies that assumption.

4.0 The minimization of causal anomalies

We haven't yet refuted Hume's argument—we've only taken the first step towards doing so. Hume could defend his view against what we've said thus far by saying the following:

> (HR)[261] Suppose that, to explain why all phi's thus far known are psi's, you posit some underlying structure or law that disposes phi's to be psi's. Unless you think that nature is uniform, you have no right to expect that connection to *continue* to hold. But if, in order to deal with this, you suppose that nature is uniform, then you're caught in the vicious circle that I described.

HR is correct. One is indeed caught in a vicious circle if, in order to show the legitimacy of inductive inference, one assumes UP; and the reason is that, just as Hume says, UP can be known, if at all, only on inductive grounds.

4.1. Why inductive inferences don't presuppose the truth of the uniformity principle

But in making an inductive inference, one doesn't assume UP and, moreover, one doesn't assume *anything* that, like UP, can be known only on inductive grounds. What one assumes is that explanations are supposed to eliminate causal anomalies—that they are supposed to reduce the number of them and to limit the scope of those that aren't eliminated. What one assumes, then, is that it is inherent in the very concept of explanation that, other things being equal, T_1 is a better explanation than T_2 if T_1 does a better job of eliminating causal anomalies than T_2.

The purpose of explanation is to minimize the breadth and depth of what must be taken for granted. The more a proposed theory requires you to say: "things just happen that way; there's no explaining it," the less successful an explanation it is.

Here's an illustration. On Monday night, you park your car in the usual place, viz., right in front of your house, which is in a quiet residential neighborhood. As usual, you make sure that you lock the car and turn the car alarm on. You also put an almost, but not quite, indestructible device (popularly known as "The Club") on the steering-wheel that locks it into place, making the car undriveable. Given where your home is in relation to where the car is parked, you'd almost certainly hear the car alarm go off, were it to do so. In fact, unless you were fast asleep, you'd even hear the footsteps of someone approaching the car. So even if the car alarm weren't on, you might well hear the approach of a would-be thief and you'd certainly hear the noises made by somebody who, not having the keys needed to open the car or to unlock the steering wheel or to start the car, would have to jimmy the door lock, saw through the wheel lock, etc.

As it happens, you hear nothing all night. And you know that you didn't sleep any more deeply than usual—that, for example, you weren't in a particularly deep drug-induced slumber—and, in general, that you weren't made less sensitive to noise than you usually are. (For argumentative purposes, we'll set aside the question of how you know this.) Nonetheless, on Tuesday morning, the car is gone.

What are the various possible explanations of this? (By "possible," I don't necessarily mean "worthy of consideration," only "not ruled out by the laws of logic.") There are infinitely many; but here are three.

(E_1) Using his or her super powers to bypass the security measures you took, some superhero-like creature made off with the car, making little or no noise and leaving few or no clues.

(E_2) As you were sleeping, the laws of physics changed in such a way that hitherto noisy processes (e.g., those that occur when car alarms go off) are no longer noisy.

(E_3) Your good friend Larry—who, despite his numerous convictions for auto-theft, simply radiates trustworthiness and decency, and to whom, without so much as a touch of anxiety, you therefore gave copies of all your keys—made off with the car. (For argumentative purposes, let's assume that, given that the car is gone, it had to have been stolen; it could not, we will suppose, have been towed or simply borrowed.)

Each of E_1 and E_2 replaces the mystery of how your car could be noiselessly stolen with a far greater mystery. In fact, each creates a whole network of deep mysteries. E_1, if correct, would generate a number of thorny questions concerning specific matters of fact and also concerning our knowledge of biological and physical law. (The features that we'd have to ascribe to the superhero in question would be ones that, unless modern biology is very wrong, a life form is most unlikely to have.) And E_2, if correct, would do nothing less than create the greatest scientific mystery of all time.

E_3 is not free from sin either. In general, you are a good judge of character. In any case, that's what you believe and that's what others tell you; and your judgments about people have generally been borne out by subsequent events. You've never been as convinced of anyone's integrity as you are of Larry's. Without being morose or censorious, he exudes a decency and centeredness that you've never even seen in a head of state, let alone a car thief. And although Larry's extensive criminal background is consistent with his stealing your car, it is not, you are convinced, *as* consistent with it as his current excellence of character is *in*consistent with it. But, of course, E_3 is

hands down the best explanation. You may think you're a good judge of character; but that doesn't mean you *are* one. People proverbially overestimate their ability to read others. And even if you are a good judge of character, so what? Anybody can be fooled; everybody *is* fooled, at some point or other. And if Larry is a career criminal—which, given his rap-sheet, he may well be—then he has probably honed his skills as a con-man.

This doesn't mean that E_3 is completely innocuous. If correct, it raises the question of how somebody could seem so different from how they are. It also raises the question of how, despite being a crime expert, you could have been unwary enough to let yourself be seduced into trusting somebody who, given his background, was so obviously unworthy of trust. But these mysteries can be dealt with more easily than the mysteries created by E_1–E_2.

All of this goes to show that *ceteris paribus* the best explanation is the one that leaves us with the fewest unexplained explainers.

Of course, new theories raise questions of their own. For example: According to Freudian psychology, we repress distressing conceits. This raises the question: "Why don't holocaust survivors repress their memories of the horrors they went through?" But Freud didn't have to concoct an *ad hoc* answer to this question, since his theory of repression is itself an answer to it. According to that theory, one doesn't repress just *any* unpleasant conceit; one represses a conceit that is in conflict with values that one has that one believes to be integral to one's psychological architecture. (One represses one's hatred of one's father because one also loves him *and*, in addition, one's life is structured around one's love for him, whereas, unless one is willing to overhaul the very structure of one's existence, one's hate for him is incapable of being acted on in a productive manner and, were it to remain conscious, would merely be a source of annoyance.) Data that seems to undermine a good theory turns out to support it when either (i) new data comes in or (ii) when what was previously implicit in that theory is made explicit. (Example of (i): Given what it known about the solar system, the trajectory of planet x is not what a certain theory says that it should be. But the discovery of a new planet y shows that x's trajectory *is* what the theory says it should be. Example of (ii): According to a certain theory, motion is relative and there is thus no difference between (a) my running around the bucket and the bucket's remaining station and (b) my remaining stationary and the bucket's rotating. But in situations of the first kind, but not the second, I become winded; and in situation of the second kind, but not the

first, the surface of the water in the bucket comes to the have the shape of the interior of a bowl. This seems to suggest that motion is not relative. But when it's taken into account that, if motion is in fact relative, one must take into account, not just my spatial relation to the bucket, but also my spatial relation to the surface of the Earth, these changes prove to be logical consequences of the theory.[262]) In any case, what's important in this context is that good theories tend get rid of causal gaps and they shrink the ones they can't get rid of.

Before proceeding, we must make a certain distinction very clear. Little or nothing about the *actual* causal structure of the world is analytic. It isn't analytic that there aren't giant causal gaps in the causal structure of the world; that reality doesn't consist of one *ex nihilo* event after another; that superheroes don't suddenly pop into existence, steal cars, and then, without leaving a trace, vanish never to be heard from again. These things, if known, are known only on the basis of sense-experience. But it *is* analytic that, other things equal, good explanations are less dependent for their truth than bad ones on the existence of unexplained explainers. Given that E_3 depends less for its truth on unexplained explainers than does either E_1 or E_2, it is *analytic* that, relative to the available data, E_3 is the better explanation.

But it isn't analytic that E_3 is *correct.* It's a possibility—maybe only a bare theoretical possibility, but a possibility no less—that the laws of physics changed overnight, that CIA super-agents have targeted you, that comic book characters now walk the Earth, etc. Anything's possible. And so far as we know that E_1–E_2 are wrong, it's through sense-perception. So, I repeat, it isn't analytic that E_3 is correct. But, given that it requires fewer unexplained explainers than others, it *is* analytic that *ceteris paribus* it's the better explanation.

4.2 Correct explanations can be bad (or, rather, badly arrived at) explanations

For an explanation to be *good* isn't for it to be correct. Sometimes the right explanations are bad ones. A story will make this clear. I'm on a bus. The bus driver is smiling. A mystery! "What on Earth does he have to smile about?" I ask myself. His job is so boring, and his life must therefore be such a horror." But then I remember that, just a minute ago, a disembarking passenger gave him fifty $100 bills as a tip. So I have my explanation: "he just came into a lot of money."

But here is the very different explanation tendered by my seatmate Gus, who, in addition to being unintelligent, is also completely insane. "The bus-driver is a CIA assassin. This morning he killed somebody who, by coincidence, had the name 'Benjamin Franklin.' Benjamin Franklin (the statesman, not the murder victim) is on the $100 bill. So when the bus driver saw those bills, he immediately thought of that morning's murder. The murder was a particularly enjoyable one; the bus driver had a chance to try out some innovative techniques. So, on being given the fifty $100 bills, the bus driver was reminded of that happy experience." *That*, my seat-mate believes, is why the bus driver is smiling. The bus driver is indifferent to the fact that he is now richer by $5,000 than he was before. (He's not a materialistic man. Plus, the CIA pays him a huge salary. $5,000 is nothing to him.)

Gus and I have access to the same empirical data. (Gus hasn't read the bus driver's diary; he doesn't know the bus driver any more intimately than I do; and so on.) And Gus doesn't have some sort of psychic gift that would give him access to otherwise unknowable facts about the bus driver's mind that would legitimize his explanation. Indeed, a belief that he has such a gift isn't even among Gus's many delusions. Nor does he have any good reason—even if, for argument's sake, one allows his delusions and hallucinations to count as good reasons—to believe that he has such a gift. So given the entirety of the data at our disposal, Gus's explanation is a bad one. The *only* evidence in favor of Gus's theory is that the bus driver started smiling upon being given the money.

But Gus is right. His explanation is correct down to the last detail.

Given that it turned out to be correct, should we say that, despite first appearances, Gus's explanation is not a bad one? No! It's a datum that it's

bad. It’s a bad explanation that turned out to be correct. Thus, for an explanation to be a good one isn’t for it to be correct.

4.2.1 An objection to our analysis

But a short extension of our story *seems* to illustrate the even stronger principle that for an explanation to be a good one isn't even for it to be *likely* to be correct.

> (NL)[263] Let *W* be the real world and let *W** be a hypothetical world where we have the exact same experiences that we have in *W*, but in which the Gus's of the world are always right.[264] In *W**, it's because Santa Claus is personally delivering gifts to millions of people that they suddenly appear in living rooms round the globe; it isn't because hard-working parents purchased them. In *W**, your keys are missing because stealthy gnomes took them; it isn't because you drunkenly threw them in the incinerator. And so on. In *W**, the right explanations are the ones that, in our world, are extremely bad ones. The data on the basis of which those hypotheses are held in *W** (so far as those hypotheses are held) is identical with the data on the basis of which they're held here (same qualification). But in *W** they're right, and the good explanations—the logical, sane, well-supported explanations—that they compete with are always wrong. So in *W** good explanations aren't even *likely* to lead to the truth. So there's no inherent connection between an explanation's being good and its being true: goodness (in the explanatory sense) cannot be identified with, or even understood in terms of, truth-conduciveness.[265]

Causation and continuity are indistinguishable. (The most obvious, and the most important, sort of continuity is spatiotemporal continuity.) Continuous sequences of events (e.g., the sequence of events composing a single light-beam) are *ipso facto* a causal process. And where there is direct causation, there is *ipso facto* spatiotemporal adjacency. (That is why Newton rightly considered the idea of "action at a distance" to be an abomination.) Discontinuities constitute partial breakdowns of an otherwise unified space-time into separate space-times. Given two events, if neither can affect the other, or if there is no third event that both can affect or be affected by, it

makes no sense to say that both events belong to the same space-time. So a "world" in which Gus is right isn't really *a* world at all: it is a collection of world-fragments. By the same token, to the extent that the data that one is trying to model represents a single world, it must be possible to show that continuous structures mediate between those data. So it *is* analytic, I would contend, that, given a *single* world, the Gus's in that world are likely to be wrong. (A "single" world, as I am using this term, is to be contrasted with one that, because of discontinuities in it, and because of concomitant fissures in its causal structures, is, as it were, powdering into more than one world.) For, if correct, the hypotheses put forth by the Gus's of a given world have the consequence that the data being modeled isn't to be understood in terms of the structure of a single, unified domain and thus isn't really capable of being explained at all---at least not in terms of any one set of principles.[266]

4.3 Explanatory goodness is a relationship, not between explanations and reality, but between explanations and data

The relations that hold in *W** between theory and data coincide with those holding in *W*. That's why the ones that are wrong in *W** are no less good on that account. What is it in W* that makes wrong explanations be good? In general, what is the relationship that good theories bear with respect to the data?

First of all, "good" is a relative to term. An explanation is good *relative* to some body of data. No good explanation is inconsistent with the relevant data. And, of the various explanations consistent with a given body of data, the best one is the one that depends the least on unexplained explainers.

Here's an illustration. I put an opaque plate on a table. I thus can't see the part of the table covered by the plate. "Given only this data," says my sophist friend Smith, "it's a theoretical possibility that the moment you put the plate on the table, there suddenly came into existence a hole in the table the exact size of the plate." Supposing that you're right, I ask Smith, why didn't the plate fall through the table? "Because," Smith responds "when that hole came into existence, a special force was created that kept the plate in exactly the place that it would have occupied if there were no hole." Why, I ask in response, did that force come into existence? "I don't know," says Smith. "It just did."

Maybe this explanation is correct. We'll never know for certain that it isn't—at least not without presupposing the falsehood of various other, comparably *ad hoc* explanations.

Be that as it may, it's a bad explanation. Why is it bad? Because it leaves us with more questions than answers. Why did the hole appear? Why did it happen to have *just* the right shape and size? Is it because the plate has special chemical properties? If so, why is it that, the moment the plate is whisked away from the table, the effects of any chemical interaction between the two are undone? Supposing that these questions can be answered, why did some force suddenly step into the breach left by the momentarily absent table surface? Why was that force *just* right—*just* enough to keep the plate on the table, but not so much that the plate rose above the table or trembled or did anything else inconsistent with its being on an ordinary table? And supposing that we have an answer to *that* question, why was there no independent evidence of the existence of that force? Why was its *only* effect

to keep the plate in that very location, for that exact amount of time? Supposing

that, for some reason or other, we have highly sensitive instruments of measurement underneath the table, why didn't any of them detect anything?

With every answer, a new hole arises. It may be that, for each hole that pops up, there is a way to plug it up that is not only consistent with the laws of logic but is *also* consistent with our other empirical beliefs. Or, if there should arise a case where the hole cannot be plugged up without jettisoning some, or even all, of our remaining beliefs, maybe there is a way to revise those other beliefs in such a way that they remain consistent with the raw data *and* with the stopgap in question.

It is generally said that a given body of data can be modeled by many different, mutually inconsistent hypotheses.[267] But even if this is true, which we'll soon see that it isn't, it doesn't follow that, for each hypothesis that models a given body of data in a way that has a given degree of success in the way of eliminating unexplained explainers, there is a logically incompatible way of modeling the data that also has that degree of success in that respect.

4.4 Hypotheses must be capable of being evaluated in terms of benchmarks that can be identified independently of the realities those very hypotheses are meant to identity

A continuation of our plate story will make it clear what this means and why it is true. Why doesn't the plate fall through the table? Because, says the conventional wisdom, the table didn't vanish. But, the skeptic will retort, maybe we can jettison that explanation; maybe there's an alternative to it that is equally compatible with the perceptual data.

But, I would respond, supposing for argument's sake that there is such an alternative, it doesn't follow that it will create fewer holes to plug up than its predecessor did. It doesn't follow that it's comparable from a cost-benefit standpoint to the old one—that, in other words, it creates as few *new* mysteries as the old hypothesis while also eliminating as many existing ones.

But not only does this not follow: the exact opposite holds. Consider the alternative explanation that we considered a moment ago (viz., *when the plate is placed on the table, a hole opens up,* etc.) The explanatory cost of that hypothesis (in other words, what it leaves us having to explain that we didn't have to explain before) was a thousand times greater than its explanatory value (in other words, what no longer has to be explained that did have to be explained before).

It would be very hard to produce an explanation that was incompatible with the conventional one and, while remaining consistent with the relevant data, was as inexpensive from an explanatory viewpoint as the conventional one. I invite you to verify this for yourself. If there is *any* reason to believe that there exists so much as a single such hypothesis, it isn't that such a hypothesis has ever been produced or ever will be. That reason, if it exists, must be purely a priori. And, for reasons that will presently become clear, there can be no a priori reason to accept such a hypothesis.

4.5 Correct explanations can be bad (not to say false) explanations (revisited)

But there is an even more immediate problem for this (or, by an analogous argument, any other non-standard) substitute for the conventional view that the table, even when obscured by the plate, continues to support it. It's far from clear that, once *all* the relevant data is taken into account, that hypothesis is even *logically* possible. Let's look at the statement:

> (*) "when the plate is placed on the table, a hole that coincides with the under-surface of the plate opens up. But a force field is suddenly created that keeps the plate exactly where it would have been if, other things being equal, that hole hadn't opened up."

Taken by itself, (*) describes a situation that could hold. That statement is a coherent one. It's not in the same category as "squares are round."

But, while that's obviously *necessary* for its being a logically feasible way of modeling the relevant data, it's far from sufficient. Think of the various physical laws that must be broken for the table section underneath the plate to vanish. Violations of natural laws, were they possible, would not be innocuous; there would be ripple effects. Think of the forces that (so we believe) *actually* prevent that plate-sized chunk of table from vanishing. Think of the various equilibriums that would be altered if those forces were immediately suspended. Think of the various energy transfers that would be changed or destroyed if, all of a sudden, those forces were put out of commission.

Those forces don't *only* keep that chunk of table in place. Were they all to be suspended, there would be many consequences *other* than the disappearance of the requisite table chunk. In other words, there would be a lot of explanatory "collateral damage," meaning that many events would occur *other* than the ones that, given that our objective is to validate (*), we want to occur. And there's no guarantee that those other events would be consistent—even in the narrowest, most emptily logical sense—with the data that, supposedly, is modeled just as well by (*) as it is by the conventional explanation (viz., "the place doesn't fall because the table chunk doesn't disappear") that it replaces.

4.5.1 Correct explanations can be bad explanations (continued)

We might try to deal with this batch of problems by positing *second-order* suspensions of natural law. So in addition to suspending those laws involved in keeping the table chunk in place, we *also* suspend those laws that, given the first suspension, lead to the just-mentioned unwanted collateral events.

The problem is that those second-order suspensions of law will be no more innocuous than their first-order counterparts. To counteract the unwanted effects of suspending the first set of forces, we have to suspend another set of forces. But just as the original forces did a lot more than keep the table intact, so these new forces will do a lot more than suspend the unwanted effects of suspending the first set of forces. And it's obviously of no use to try to deal with this problem by bringing in a third set of forces.

"You may be right," it will be said, "that *if* we hold onto our existing beliefs about physical law, *then* the disappearance of the table chunk would have all of this unwanted fall-out. But surely we could adjust those beliefs in such a way as to prevent this. Whatever those beliefs are, just get rid of them and replace them with ones that *don't* lead to all of this unwanted fall-out."

The very fact that this move must be made only proves my point, namely, that (*) is *not* consistent with what we know. In fact, it proves a stronger point, namely: (*) isn't consistent with *any* body of beliefs *at all* like the ones we now have.

The objector is saying, rightly, that there is *no* way to square (*) with our existing beliefs—that it is only what we believe *minus many of our beliefs about natural law* that can be so squared. But when our beliefs about natural law are subtracted, so to speak, from the totality of our beliefs, the resulting body of beliefs bears very little resemblance to our current beliefs. Our beliefs about natural law are not in all cases neatly separable from our beliefs about specific matters of fact; and if we were to strip away enough of the former to compatibilize (*) with our belief system, we'd be biting off a lot more than we bargained for. This is because various beliefs about natural law are embedded in the concepts associated with expressions that, in order to say what's on our minds, we must use every day (e.g., "microscope," "telephone," "radio," "solar panel," "cell phone," "thermometer," "thermostat"). Theory-laden though they are, we need those terms to describe facts of everyday, pre-theoretic experience. We therefore need them to state the pre-theoretic data that our theories are to model.

“But,” it will be said, “anything that can be described *only* by taking for granted some theoretical belief is *ipso facto* not a *datum.* A truth? Maybe. A datum? No.”

That may well be true. But, if so, our belief-system is so overrun with theoretical beliefs that no belief system free of such beliefs would bear even the slightest resemblance to it. Also, it is incoherent to suppose that there could exist a belief-*system* that wasn’t replete with theoretical content. The difference between a mere belief, on the one hand, and a belief-system, on the other, is that belief-systems are intended to be explanatory, whereas mere beliefs needn’t have this property. I believe that it’s raining because I see that it’s raining—*not* because, in an effort to model various observational data, I at some point *hypothesized* that it was raining. But psychoanalytic theory constitutes a belief-system, as opposed to a mere belief, because it is intended to be explanatory. This suggests (though it doesn’t establish) that “belief-system” and “theory” are veritable synonyms. (Why the parenthetical hedge? “Because there are belief-systems that aren’t of a theoretical nature. Religious belief-systems aren’t theories.” But even this is arguable.[268] Some would say, and I agree with them, that religious systems are proto-theories or para-theories—structures that make heavy allowances to emotional factors and make insufficient allowances to considerations of logic and evidence, but are, within these parameters, as theory-like as they could possibly be.) If, indeed, belief-systems are *ipso facto* theories, it makes no sense to speak of what our “belief-system” would be like if it were stripped of theoretical content. Long story short: although it’s obvious that, on its own, (*) is perfectly coherent, it’s far from obvious that (*) is compatible with our belief system as a whole or with any belief system that isn’t radically different from it. This was the very point we set out to establish. It goes without saying that what is true of (*) holds *mutatis mutandis* of each of many other “deviant” hypotheses.

But even if, for some reason, this last point is denied, there is no way to deny our main point, to wit: even if (*) is logically consistent with our belief system, or at least with some not-too-mutilated version of it, the explanatory benefits of countenancing (*) in favor of the conventional explanation are grossly outweighed by the explanatory costs.

4.6 Synthesizing these points

It will help if, before moving forward, we take a moment to synthesize the points thus far made. There's no difference between committing the Gambler's fallacy and performing an induction by enumeration. At the same time, knowledge of invariable concomitances is obviously an extremely valuable source of knowledge. But it isn't, or in any case shouldn't be, through enumerative induction that one transitions from (1) "all phi's that I've known of were psi's" to (2) "the next phi I come across will be a psi." (In this context, it is to be assumed that that "*x* is a phi" doesn't analytically entail "*x* is a psi.")

The *actual* bridge between (1) and (2), supposing that there is one, lies in the fact that, given (1), it is reasonable to believe there to be some mechanism or natural law that requires (or at least disposes) phi's to be psi's. By itself, (1*) "all metal known to me has expanded when heated" provides no good reason to accept (2*) "all metal expands when heated." (1*) provides a good reason for accepting (2) *only* to the extent that (1*) provides a good reason for accepting (3*) "there is a lawful connection between a thing's being made of metal, on the one hand, and it's expanding when being heated, on the other."

4.6.1 Synthesizing these points (continued)

Before we continue, it's crucial to note that none of this begs any questions against Hume. Hume is saying: "given only that every metal object that Smith *has* known of has expanded when heated, it doesn't follow that metal objects that Smith has *yet* to encounter will (or is even likely to) expand when heated." We agree. Smith is guilty of the gambler's fallacy so far as he thinks otherwise. All we're saying is this: Smith's experience thus far *suggests* that "*x*'s being heated metal provides some kind of nomic basis for *x*'s expanding" *has* held for objects falling within the scope of Smith's experience.

We're *not* making the claim Smith's experience *entails* that some such connection has held. Such a claim would simply be false. And we're *not* making the claim that Smith's past experience gives him a good reason to think that metal objects not (yet) known to him will be more likely to expand or even that they'll be more likely than not to expand. Such a claim would beg the question against Hume. Finally, we're *not* making the claim that Smith's past experience gives him a good reason to think that there *continues* to exist any nomic connection between a thing's being heated metal and its expanding. Echoing what we just said, such a claim would beg the question against Hume: for Hume's very point is that a thing's happening in the past is no guarantee of its continuing to happen in the future. And we cannot, without begging the question against Hume, assert that a law's having held in the past is evidence of its holding in the future.

So what *are* we saying? Only that Smith's experience gives him a reason to believe that such a connection *has* held, at least as far as metal objects falling within the scope of experience are concerned. And in making this obviously plausible claim, no questions have been begged against Hume.

4.6.2 Synthesizing these points (continued)

I propose that, armed only with this non-question-begging point along with the tautology that it is the purpose of explanations to eliminate unexplained explainers, we can show that Smith has reason to expect *future* metal objects to expand when heated.

We've agreed that Smith's experience does give him some reason to think that, at least as far as objects thus far known to him are concerned, a thing's being metal provides some kind of causal or nomic basis for its expanding when heated. Let *P* be this causal principle. *P* is of doubly limited scope. It applies only to *past* objects *and* it applies only to past objects that Smith knows of. Let's suppose that Smith accepts *P*.

Smith has two options. (I am not counting a refusal draw any inferences as an option.)

> (i) Hold that *P* provides no reason to believe that, outside Smith's past, personal experience, a thing's being a metal provides a basis of some kind for its expanding when heated.
>
> (ii) Hold that *P* does warrant some such generalization.

One such generalization is: "that connection holds for all past objects, whether Smith knows of them or not; but it doesn't hold of future objects." Another is: "that connection holds for objects that Smith *will* know of, but not for those that he won't know of." But, of course, the most natural, not to say the most rational, generalization is this:

> (*P**) "That connection holds *generally*. It holds no less for objects not known by Smith than for those known by him, no less for future objects than for past."

Let's suppose that Smith chooses *not* to accept *P** or any other generalization of *P*. How, in that case, would the world have to be for Smith's belief system to hold? Metal objects that Smith became aware of would suddenly undergo a profound change: they would go from being things that, when heated, didn't expand to being ones that did. And for this change to occur, each atom of every metal object known to Smith would have to

change in fundamental ways. Given an acceptance of *P*, it may seem conservative not to accept any generalization of *P*. But it's the very opposite. If you accept P but reject any generalization thereof, you're saying that objects miraculously undergo profound microstructural changes whenever you know of them and undergo equally profound changes when you stop being aware of them. And *that* hypothesis is not a conservative one.

And, quite clearly, it's a bad one. Why is it bad? Because it creates more unexplained explainers than it eliminates—because it creates more questions than it answers.

4.6.3 Synthesizing these points (continued)

Of course, *not* accepting a hypothesis isn't the same as rejecting it. *Not* accepting some generalization of *P* isn't the same as rejecting it. And it's obviously a good idea in many cases to hedge—to wait for evidence. But that's irrelevant. The question is: given that there is enough information to warrant an acceptance of *P*, is there enough information to warrant acceptance of any generalization of *P*? If we answer by saying "no," then we're saying that, given only the evidence at our disposal, it's no less likely than not that objects systematically undergo massive changes at the atomic level whenever they come or cease to be known by Smith. And *that* position, as we saw, creates more mysteries than it eliminates. So despite first appearances, there's nothing conservative about refusing to countenance any generalization of *P*. And such refusal embodies acceptance of a principle that is straightforwardly false, namely: "there's no reason to prefer explanations that explain a lot and create little that has to be explained to those that explain a little and create a lot to be explained."

4.6.4 Synthesizing these points (continued)

These points are easily generalized. It's no less absurd to suppose that, in the last nanosecond, all metal objects have undergone profound microstructural changes than it is to suppose that all metal objects known by Smith are microstructurally very different from those not known by him. This point provides the non-circular justification for UP that Hume sought but couldn't find.

Let's suppose that Smith has reason to believe (i) that, where metal objects that he *has* known of are concerned, a thing's being heated provides some kind of basis for its expanding. Let's also suppose that Smith holds (ii) that this principle *won't* hold for metal objects that he *will* encounter (i.e., for metal objects that he hasn't encountered but will and for those that he has already encountered and will encounter again). In holding (ii), he is *ipso facto* supposing that the laws of nature will abruptly change. And in positing this one change, he is positing innumerably many others, as we will now see.

A metal object's being heated involves innumerable microstructural changes, and so does its expanding. The macroscopic supervenes on the microscopic. Given two objects that are microstructurally alike—that consist of the same particles interrelated in the same way—there is no possibility, even a purely logical one, that those two objects should differ at the macroscopic level. By the same token, given two objects that differ at the macroscopic level, there is no possibility, even a purely logical one, that those two objects should *fail* to be microstructurally different. A metal object's being heated thus consists in its undergoing many microstructural changes, and so does its expanding. And if a metal object's being heated causes it to expand, it is in virtue of the fact that the first set of microstructural disturbances result in, or are themselves constitutive of, the second set.

So given a law—even one that was confined to objects that Smith has encountered—to the effect that metal expands when heated, in order for that law to be annulled, each of innumerably many microparticles would suddenly have to undergo quite dramatic changes. So in assuming that, all of a sudden, a law that once held ceases to do so—or, what is the same thing, in assuming that the future, starting now, is governed by different laws from the past—one is positing not *one* change, but innumerably many. For, in making that assumption, one is (implicitly) supposing that each of the innumerably many

microparticles will, all of a sudden, stop being governed by one set of laws and will start being governed by another. And one is thus creating what, for practical purposes, might as well be infinitely many unexplained explainers.

But not a single one of these unexplained explainers is created by the hypothesis that metal objects that we will encounter are governed by the same laws as those we have encountered. So given the truism that one explanation is better than another if, other things being equal, the first creates fewer unexplained explainers than the second, it follows that the rational move, at least until further evidence comes in, is to suppose that those microparticles will continue to be governed by the same laws that used to govern them.

According to what is now the conventional wisdom, which coincides with the viewpoint that Hume bequeathed to us, the conservative move is the agnostic one: stick to the evidence; generalize as little as possible; don't assume that what you don't know resembles what you do know. But the so-called conservative view is really *less* conservative than the alternative. (The alternative is: "assume that what you don't know *does* resemble what you do know.") The so-called conservative view posits giant schisms in the natural order and, therewith, masses of unexplained explainers. The alternative view does not.

5.0 The goodness of hypothesis a function of its relationship to the data, not to the truth

Earlier I said:

> (i) that explanatory goodness is a relationship between a given hypothesis and the available data.

And I explicitly *denied*:

> (ii) that explanatory goodness is a relationship between a given hypothesis and *the truth.*

"But surely this makes no sense," it will be said. "Hypotheses are attempts to identify the truth. So surely for a hypothesis to be good is for it to be true; and for one hypothesis to be better than another, is for it to be closer to the truth than the other."

Here's the problem with this reasoning. Setting aside questions that can be answered *directly*—in other words, setting aside questions (if any there be) for which, without making any inferences, one can produce justified answers---it is through hypothesis, and *only* through hypothesis, that we have any way of identifying the truth. Often, when trying to solve a particular problem, we have many reasonable hypotheses to choose from. When seeking the truth, one must choose the *best* hypothesis available. This is a tautology. If, for any hypothesis H, H's being good was the same as its being true, we'd have no grounds for accepting Hypothesis 1 as opposed to Hypothesis 2 unless we *already* knew what the truth was---unless, in other words, we *already* knew the very things of which it was the purpose of those hypotheses to apprise us. If hypothesis H's being good were identical with its being true, then we'd have no way of knowing whether H was good *until* we knew whether it was true. But if we didn't know whether a given hypothesis was good until we knew whether it was correct, hypotheses would be useless in the way of helping us *find* the truth. For, by supposition, they couldn't be evaluated *until* their job had already been done.

It follows that, when choosing between hypotheses, our choice cannot, if justified, be based on how *accurate* those hypotheses are. For if we knew how accurate they were, we wouldn't need them to find the truth and we thus

wouldn’t need them at all. Thus, when choosing between hypotheses, our choice (supposing it principled) must be based on something that, unlike the truth, is available to us. In other words, when trying to figure out to what extent a given hypothesis is worthy of our acceptance, the yardstick with respect to which we compare that hypothesis cannot possibly be the truth. For obvious reasons, that yardstick cannot be anything other than the relevant data.

Let us sum up. The degree to which a given hypothesis is good is obviously something we consider when determining whether or not to accept it. The purpose of that hypothesis is to help us identify some as of yet *unknown* truth. Therefore, the degree to which a hypothesis is good is *not* identical with the degree to which it conforms to the truth. Therefore, the degree to which a hypothesis is good is identical with the extent to which it makes sense of the data. Therefore, explanatory goodness (i.e., the sort of goodness had by hypotheses *qua* hypotheses) is a relation between hypothesis and data, not between hypothesis and truth.

A corollary is the point (which we stressed but which some philosophers of science deny[269]) that *bad* hypotheses can be correct ones. For the degree of goodness of a hypothesis, being the function of its fidelity to the data, not to the truth, a totally false hypothesis can be good and a completely accurate hypothesis can be bad; for the false one might model the available data, and thus satisfy the relevant desideratum, whereas a true one might fail to satisfy that desideratum.[270]

6.0 Closing the argument

We can now succinctly say why Hume's argument fails.

> Step 1: It's an *analytic* truth that *ceteris paribus* the better explanation is the one that creates fewer unexplained explainers. Let's refer to this principle as "MC" (short for "minimization of causal anomalies"). If UP can be known, it must be known on inductive grounds. But MC is not to be known on such grounds, since it's analytic and, unlike UP, is to be known on purely conceptual grounds. (It makes no sense to say that theory T_1 is better than otherwise comparable theory T_2 if T_1 posits more causal anomalies than T_2. It makes no sense to say that T_1 posits fewer causal anomalies than T_2 if T_1 posits more unexplained changes than T_2. It makes no sense to say that T_1 posits fewer unexplained changes than T_2 if T_1 posits more unexplained breaches with the past than T_2. And it makes no sense to say that T_1 posits fewer unexplained breaches with the past than T_2 if T_1 posits more spatiotemporal discontinuities than T_2.) Thus, MC, unlike UP, can be non-circularly assumed in the context of a defense of inductive inference.
>
> Step 2: It's a datum that various concomitances are to be found *among the objects known to one.* It's a datum that, so far as the object's known to (say) *Smith* are concerned, putting a hand in boiling water causes intense pain, whereas putting that same hand in cool water causes intense relief.
>
> Step 3: One creates more unexplained explainers by assuming that the unknown differs from the known than by assuming that it does not.
>
> Step 4: If one remains agnostic—if one refuses to say, one way or the other, whether the unknown resembles the known—

one is committed to the incoherent position that there's no reason to prefer a hypothesis that creates few unexplained explainers to an otherwise comparable one that creates many. (The reason this view is incoherent is that for one explanation to be preferable to another just *is* for it to do more in the way of explaining, and thus less in the way of creating unexplained explainers, than the other.)

Step 5: Given MC, and given Steps 1–4, it follows that there is an analytic and therefore *non*-inductive reason to hold that, other things being equal, the unknown resembles the known.

Conclusion: Hume is wrong. This is a direct consequence of Step 5. Contrary to what he says, it *is* rational to hold that, in its more general respects, the known is a good indication as to what the unknown, in *its* more general respects, will turn out to be like, supposing that we come to know it. Given the legitimacy of MC, it is obvious why skepticism about the external world is actually an extravagant and irrational position, though it may initially appear cautious and conservative. One's sensory experience is replete with discontinuities. Every time you open your eyes or turn your head, you are flooded with sensory experiences that are totally unlike those that you were experiencing a moment earlier. According to the non-skeptic, these new experiences don't come into existence *ex nihilo;* they are merely links in chains of events whose other links you don't perceive. The skeptic says that it's no more rational to believe this about our sensory experiences than it is to believe that those experiences come into existence *ex nihilo.* In saying this, the skeptic is saying, in effect, that it's just as rational to accept a theory that creates more anomalies than it gets rid of than it is to accept an otherwise comparable theory that gets rid of more anomalies than it creates. Since it's an analytic truth that, given two otherwise comparable theories, the one that leaves us with fewer anomalies is the one that ought to be preferred, it follows that the skeptic's position is analytically false. Of course, it

isn’t analytic that our sensory experiences are correct. But it *is* analytic that it’s rational to suppose our sensory experiences veridical *to the extent that doing so* leaves us with fewer causal anomalies than we’re otherwise left with. The skeptic’s denial of this embodies an analytically false conception of explanation.

6.0 Skepticism about the external world

Given what we've said about induction, it is clear why, on the whole, it is rational to trust one's senses---why, in other words, skepticism about the external world is not a rational view. There are two kinds of perceptions: those that have mental predecessors that they resemble and those that don't. If, while standing still and otherwise undergoing no changes relevant to your visual experience, you stare at a blank wall to which nothing is happening, the visual experience you have at any one instant will be very similar to its immediate predecessor. But, typically, there are discontinuities in one's visual experience. The visual experience upon opening your eyes bears *no* resemblance to the visual experience you had before you opened your eyes. First, so your visual experience tells you, there is nothing. Then, so your visual experience tells you, there is a whole world of objects. And, supposing your visual experiences to be veridical, the states of affairs that you see after blinking are very similar to those that you saw just before blinking.

Looked at from the skeptic's perception, there are three possibilities as to what is going on when a person blinks:

> (a) During the instant that his eyes are shut, the situation he saw vanishes and, when he opens his eyes, there *ex nihilo* pops into existence a situation a numerically distinct from that situation but similar to it, if not qualitatively identical with it.
>
> (b) Before the person blinks, he isn't actually seeing anything. He is hallucinating. During the moment that his eyes are shut, he stops hallucinating. When he opens his eyes, there *ex nihilo* pops into existence a hallucination numerically distinct from the pre-blink hallucination but similar to it, if not qualitatively identical with it.
>
> (c) Before that person blinks, he is actually seeing various objects. His pre-blink visual experiences are caused by light-rays bouncing off of external (i.e., non-mental) objects. Information concerning those external objects and their interrelations is coded into the light rays that collide with his eyes; that structural information is transmitted to some part of

> his brain that produces visual experiences that themselves encode it. When the person's eyes close, light stops colliding with his eyes. During that brief period, nothing collides with his eyes that bears information about anything external. (Light that bears such information collides with his eyelids. But his eyelids don't in their turn transmit that information to the relevant part of his brain.) After he opens his eyes, light-rays bearing information collide with this eyes, leading by way of the previously described process to visual experiences that bear structural information concerning their external causes. So the reason that his pre- and post-blink visual experiences are so similar is that they have commensurately similar external causes.

The non-skeptic accepts (c). The skeptic neither accepts nor rejects any one of (a)-(c). The skeptic says that, given any one of those hypotheses, it is no more rational to accept it than it is to accept either of the other two. So, according to the skeptic it is more rational to reject (a)/(b) in favor of (c) than it is to reject (c) in favor of either (a)/(b).

Right now, we'll argue that (c) is preferable to (a). What we say in this connection will make it clear why (c) is preferable to (b).

If one accepts (a), one has to explain two things that one doesn't have to explain if one accepts (c). One has to explain why, the moment one opens one's eyes, *anything* comes into existence. Second, one has to explain why, despite any causal connection between the pre- and post-blink situations, the post-blink situation is so similar to the pre-blink situation. And, in having to explain each of these two things, the advocate of (a) has to explain innumerably many things that the advocate of (c) does not have to explain. The advocate of (a) has to explain the *ex nihilo* creation of each blade of grass that the person in question sees when he opens his eyes, and he must do the same for each other object post-blink object. Further, he must explain why (setting aside the odd post-blink object that doesn't have a pre-blink counterpart) every one of the objects that *ex nihilo* popped into existence is so similar, in so many respects, to one of the pre-blink objects that was annihilated. (These respects include, but are not limited to: color, shape, position, and state of motion.) If (a) is right, every one of these changes is a singularity---a breach in the fabric of space-time. (In fact, we'll see in

Chapter 24 that, if (a) is right, the pre- and post-blink worlds are *too* disconnected from each other to be considered ***genidentical***[271]-- other words, that the occupants of the post-blink world are not even in the same space-time manifold as the occupants of the pre-blink world.) But if (c) is right, neither the differences nor the similarities between the pre- and post-blink worlds constitutes a breach in the fabric of nature. The differences correspond to continuous changes, and the similarities correspond to ***diachronic*** identities and thus themselves, for reasons that we'll see in Chapter 16, correspond to continuous changes.

Chapter 12 Empiricism and Its Limits
1.0 Empiricism vs. rationalism

We obviously acquire a great deal of knowledge through "sense-perception" (i.e., through sight, hearing, touch, and so forth). According to a doctrine known as "empiricism," *all* knowledge is derived from sense-perception.

According to a view known as "rationalism," some knowledge is acquired entirely through the use of one's ability to *reason.*

Rationalists almost never hold that *no* knowledge is acquired through sense-perception. They hold only that reason, as opposed to sense-perception, is the vehicle through which *some* knowledge is acquired. Rationalists typically hold that knowledge acquired in this way is very important—it isn't trivial.

Some hold the view that there is knowledge that is acquired neither through the senses nor through reason. I don't wish to dismiss this view. Maybe it's correct. But there is an apparent problem with it. Any case of knowledge is a case of *justified* true belief. Given a belief that isn't acquired through the senses *or* through reasoning, the question arises: what could possibly justify it? And there's no obvious answer.

In any case, in this chapter, if only to simplify exposition, it will be assumed that the only viable sources of knowledge are reason and the senses, and therefore that the only viable options for epistemologists are rationalism and empiricism. But I'm inclined to think that there is. Part of my reason for believing this is that we seem to have direct awareness of certain non-empirical facts (e.g. the fact that green isn't orange)[272], and *direct* knowledge *ipso facto* is not obtained through *analysis*. Analysis is an operation that is carried out on pre-existing knowledge and, when properly executed, therefore yields *meta-knowledge* (knowledge about knowledge); and it is obvious that meta-knowledge cannot possibly be direct. We do therefore seem to have a faculty of divination, as it were, whereby we directly apprehend non-spatiotemporal truths.

Those who believe that there exist non-spatiotemporal entities are rationalists (or at least must be rationalists, lest they be guilty of inconsistency.) Give or take a few nuances, empiricism is the view that you don't know it if you don't see it. We see property-instances, but not properties themselves. Therefore, empiricists either deny the existence of

properties or they hold that properties are identical spatiotemporal entities. Those who take the latter view are forced to view properties as being composed of their own instances; so the property of being wet is identical with some object composed of all wet things, or some such. Therefore, empiricists don't believe in *uninstantiated* properties. Therefore, anyone who does believe in such properties, or who believes that properties are not spatiotemporal entities, is a rationalist.

Those who believe that properties are non-spatiotemporal are *Platonists.* Therefore Platonists are rationalists. A Platonic metaphysics requires a rationalist epistemology. Plato himself accepted a very extreme form of rationalism.

Being one that has deep roots in common sense, empiricism is a very old doctrine. But empiricism was first rigorously developed by John Locke (1632–1704), George Berkeley (1685–1753), and David Hume (1711–1776). The doctrines put forth by these authors form the basis of modern philosophy. Hume's beliefs about *causality* and *inductive inference* are outgrowths of his empiricism. These views are thoroughly examined in Chapters 12 and 17.

Berkeley's belief that objects are identical with our perceptions of them is an outgrowth of his empiricism. Berkeley's ingenious arguments for this outrageous doctrine are examined later in this chapter.

John Locke's plausible position that universals are "the workmanship of the understanding"—that *we* create them, so as to make sense of the world—is a derivative of his empiricism. That doctrine is examined in the present chapter. (It is also examined in Chapter 2.)

The first great rationalist was Plato (428 B.C.–348 B.C.). (Plato's views are defended in Chapter 2.) Other rationalists are: Augustine (354–430), Descartes (1596–1650), Leibniz (1646–1716), and Gottlob Frege (1848–1925).

The first great empiricist was Aristotle (384 B.C.–322 B.C.), whose views, along with Plato's, are also considered in Chapter 2. Other empiricists are J.S. Mill (1806–1873), Rudolph Carnap (1891–1970), and W.V.O. Quine (1908–2000).

In this chapter, we will find that empiricism is not in its strictest form a defensible doctrine and that, consequently, at least some knowledge has a non-perceptual basis.

There are, as we will see, two important corollaries. First, many commonsense-based views about the internal (psychological) and external

(physical) worlds must be jettisoned. Second, sober scientific methodology actually demands an acceptance of principles that would strike common sense as decidedly unscientific. We will see that many concepts that cannot be adequately understood along strictly empiricist lines can nonetheless be correctly understood only *in terms of* what empiricism has to say about them.

2.0 The particularist nature of sense-perception and the need for universals

Sense-perception discloses only *particulars* to us—particular objects having particular property-instances. But we understand and describe the world in terms of categories or, as they are sometimes called, "universals." These can be thought of as natures, essences, or "ways of being," that, although *exemplified* by the particulars that we encounter in sense-perception, are not themselves to be sense-perceived, and, though they presumably exist, do not exist in space-time.

Because they are non-spatiotemporal, those of an empiricist mindset regard universals with extreme suspicion. In fact, empiricism is, nearly enough, the denial that such things exist. Why? The empiricists say: what we can legitimately believe is what we can sense-perceive or know on the basis of extremely conservative inferences on the basis of what we can sense-perceive.

The problem is that there is no way to understand the world except in terms of universals. To understand things *is* to know how, in virtue of having such and such characteristics (i.e., in virtue of falling under such and such universals), they are subject to thus and such principles or laws. "Why did Max steal the money? Because Max is greedy and amoral and he believed it to be to his strategic advantage to steal the money." Max's behavior is explained by showing how, given the operative conditions, his having those characteristics makes it inevitable that he will act in that way. This explanation presupposes the existence of many different universals. (Among them are those denoted by the terms "greed," "amorality," "money," and "theft.") This explanation is not unique in this respect: *all* explanations presuppose the existence of universals. The very idea of an explanation that didn't do so is an incoherent one.

Universals are outside of space-time. For this reason, they cannot have effects on anything. (Causal relations hold only among spatiotemporal entities.) One cannot sense-perceive an object without being affected by it. (It is only because your ocular surfaces are disturbed by light-rays bouncing off of the page in front of you that you can see it.) It follows that universals cannot be sense-perceived. It also follows that no *evidence* of them can be sense-perceived. This is because *x* is evidence of *y* only if *x* and *y* are *causally*

interconnected in such a way that, given *x*'s existence, *y*'s may reasonably be assumed. (Smoke is evidence of fire because smoke and fire are causally connected in such a way that, given the presence of smoke, the presence of fire may be inferred.) It is therefore impossible as a matter of logic to find any *observational* basis for the existence of universals. This means that, if empiricism is correct, we cannot possibly have any good reason to posit the existence of universals. Given how endemic universals are to thought, language, and explanation, the right move is to reject empiricism.

2.0 The shortcomings of empiricist attempts to deal with these problems

Empiricists have responded by saying that, although *instances* of universals obviously exist, universals *per se* do not. (Remember that, in this context, "universal" is synonymous with "property.")

The problem with this view is that, as previously noted, every sentence and every thought demands the existence of at least one universal.

Also, some basic logic shows that universals exist. Consider the following inference: "Bob and Ted are both human; therefore, there is some characteristic that they have in common." That is a valid inference; and the conclusion affirms the existence of a universal. ("Characteristic" is synonymous with "universal.")

Finally, many statements straightforwardly presuppose or affirm the existence of universals; and were it not for statements of this kind, many obvious truths couldn't be expressed. Consider the following statements:

> (i) Hitler and Stalin had a lot in common;
>
> (ii) Smith has what it takes to be a great pianist, but he doesn't have what it takes to be a great composer;
>
> (iii) Even though Smith is different from Jones, that has nothing to do with the fact that they don't like each other.

The meanings of these sentences are, respectively:

> (i*) There are many properties such that, given any one of them, Hitler had it and so did Stalin;
>
> (ii*) There are properties such that, if one has them, one is a great pianist and such that, moreover, Smith has them; and there are properties such that, if one lacks them, one isn't a great composer and such that, moreover, Smith lacks them;
>
> (iii*) There are properties such that Smith has them and

> Jones does not, but this fact isn't responsible for the fact that they don't like each other.

(i)–(iii) cannot be translated into statements that don't presuppose or affirm the existence of properties. So it isn't really an option to hold that properties don't exist.

As soon as one grants that anything resembles anything, one must grant that universals exist independently of thought. For two objects to be similar in some respect is for there to exist some property that they both have. For two objects to have the same shape is for there to exist some morphological property such that each object is an instance of that property. It's obviously possible for two objects to have the same shape in a world devoid of sentient beings. Therefore, there exist properties in worlds where there are no sentient beings. Therefore, properties aren't created by such beings and aren't otherwise dependent on them.

2.1 More failings with empiricist attempts to deal with these problems

Properties are identical with *categories.* For *x* to fall into a given category is for *x* to have a given property. (For *x* to fall into the category of round things, is for *x* to have the property of being round.) Given that properties exist independently of thought, the same is true of categories. Thus, contrary to what some empiricists (e.g., Locke) say, categories are not devices that human beings created in order to expedite thought.

It is easy to find independent corroboration for this view. Let each of *x* and *y* be a pint of pure, clean glacial water, and let *z* be a pint of gasoline. The chef or chemist or auto mechanic who chooses to co-categorize *x* and *z*, as opposed to *x* and *y*, is in for some very unpleasant surprises. Why? Because his categories don't cut "nature at the joints," to use Francis Bacon's expression. In other words, there is some relevant property that *x* and *y* have in common that *x* and *z don't* have in common, and this person's system of categories embodies an ignorance of that fact.

3.0 Non-perceptual knowledge a prerequisite for perceptual knowledge

It's not as though tables, rocks, trees, etc., create little images of themselves that they deposit in our hitherto empty crania. Our minds have to supply the "paint," so to speak, with which our mental representations of those objects are painted.

In fact, our minds must supply *everything.* External objects don't *supply* our minds with anything. All they do is tell our minds when to deploy what is already in them. External objects don't stock our minds; they *activate* what's already in them. A rock will never experience an image. It doesn't matter how many light beams you fire at it. There's nothing in it to be brought forth; there are, so to speak, no images in there to begin with. For similar reasons, the acoustical disturbances that, in a human, would lead to knowledge of English or of philosophical principles will have no comparable effect on a rock. Knowledge is transmitted by being elicited, not by being deposited. A failure to see this is the most fundamental problem with empiricism.

A precondition for sense-perceiving anything is that we be able to represent it as having a location (relative to us) in space and time, that we be able to represent it as having a certain shape, state of motion, color, etc. Although we learn a great deal through perception, what we *don't* learn through it are the concepts—space, time, motion, shape, color, etc.—out of which sense-perceptions are constructed. So those things, and all that having them involves, must be in the mind *prior* to sense-perception, as a necessary *pre*condition for it.

These considerations expose what I believe to be a deep confusion of which empiricism is guilty. Empiricists fail to distinguish between two very different roles that sense-perceptions can have: they can have a *triggering* role and they can also have an *information-transmitting* role.

3.1 Sense-perception as transmitter of information about the external world

You see a dog. On that basis you have the belief that there is a nearby dog. Here your eyes transmitted information to you about the external world. They "told" you, so to speak, that there was a dog in a certain place. You could not possibly have learned this fact through pure logic-chopping or conceptual analysis. You could learn it only through the senses (either by seeing the dog or by sense-perceiving some other sort of evidence of the dog). In this context, then, information about the external world is being transmitted to you through your senses. Your senses (specifically, your sense of sight) is thus functioning in what we might call an "information transmitting" capacity.

3.2 Sense-perception as trigger

Sense-perceptions have a triggering role when they provide the recipient with new information *not* by telling him something about the external world, but by activating or "triggering" some already existing cognitive faculty.

You are in a geometry class. The teacher draws something that is supposed to represent a triangle but only vaguely looks like one. He draws various lines through it, around it, etc. And from these diagrams you learn various facts about triangles.

What did you actually *see?* What did your eyes show you? A triangle? No—only a few uneven lines that formed something that vaguely looked like a triangle. How did your seeing this non-triangle get you to know that, for example, the area of a triangle is 1/2 base times height? Since you didn't see any triangles, you didn't see any triangles' height, base, or area. The things you saw were symbols for things that had to be supplied to you by something other than your senses—that had to be supplied to you by your intellect.

This is not to say that the diagram was useless. Plainly it was not. It may be that, but for seeing it, you wouldn't have learned anything about triangles. But your seeing the diagram led to your knowing those truths about triangles *not* because any triangle, or any fact about triangles, was any part of the information imparted to you by your perception, but rather by *triggering* or *activating* certain conceptual faculties of yours. In this context, your senses are (to the extent that you are learning actual geometrical truths as opposed to mathematically irrelevant facts about the blackboard, chalk-marks, etc.) providing you with knowledge by getting you to acquire knowledge in some *non*-sensory way, viz., by getting you to grasp certain concepts to and to identify some of the relations holding among them.

The proof that your visual perception of the diagram had only a *triggering* role, and that it did not itself contain any information relating to what you learned, is that, if you were to forget ever having seen that diagram—were to forget what it looked like, who drew it, where you saw it, or even *that* you saw it—it would have no effect on the geometrical truths that you learned that day. The reason is that those geometrical truths (e.g., area of a triangles = bh/2) were simply *not* part of the content of what your senses were telling you.

By contrast, if you forget ever seeing the dog (described in the last section), you *will* forget the information that, on that occasion, your senses

led you to have. The reason is that, in that context, your senses weren't serving the purpose of jogging your ability to acquire knowledge in some *non*-sensory way (viz., through rational insight) but were themselves the bearers of the information learned.

3.3 An anti-empiricist consequence of the distinction between triggering and information transmitting

According to strict empiricists, *everything* that is learned, including purely logical principles, is learned through sense-perception. When this appears to happen, it's because sense-perception is simply "turning on" some non-sensory mode of knowledge acquisition. "Truths of reason," as Hume called them, aren't learned through sense-perception, even though sense-perception may awaken the non-sensory faculty through which we do become aware of them.

Some empiricists are aware of this objection, and they generally deal with it by saying one of two things:

> (1) Truths of reason *really are* truths about the irregular figures on blackboards, that truths of arithmetic *really are* truths about what happens when apples are put on top of other apples, etc.
>
> (2) The so-called "truths of reason" are just linguistic or typographical conventions. "1 + 1 = 2" is an empty, purely definitional truth. "2" is *defined* as "1 + 1," and "3" is *defined* as "1 + 2." Every other fact about mathematics can be reduced to definitions, linguistic conventions; and all of mathematics turns out to be typographical. (As a matter of typography, you can replace "1 + 1" with "2," just as you can replace "five" with "5.") And all the so-called "truths of reason" that lie outside of mathematics are thought to be explained in the same way.

3.3.1 Why (1) is wrong

Your knowledge that 113 + 876 = 989 isn't empirical; you didn't see 113 pears being counted, then see 876 pineapples being counted, and then see the whole lot being counted. (And even if you did have such an experience in connection with those particular numbers, there are many other pairs of numbers whose sums—not to mention quotients, products, differences, etc.—you know, or could readily figure out, but in connection with which you've never seen anything like the situation just described.) Your knowledge of arithmetical truths is based on a grasp of non-empirical principles.

Even if (what we will find to be false) you learned that 1 + 1 = 2 from watching one thing be put on top of another, you still would not, on that basis alone, have any way of generalizing that knowledge—of inferring the other mathematical truths you know. To make the needed inferences, you'd need access to rules of inference that you couldn't possibly know about if your knowledge were strictly sense-based. A rule of inference is given by a principle to the effect that, given such and such conditions, thus and such *must* be the case. Our senses don't tell us what must be the case; they tell us, at most, what *is* the case. Therefore, it cannot be strictly through sense-perception that we learn the inference-rules on the basis of which we can extract knowledge of a general and abstract kind from the, in all cases, hyper-specific information given us by our senses.

3.3.2 Another reason why (1) is wrong

Even one's knowledge that 1 + 1 = 2 isn't empirical. As David Hume made clear, observation tells you how things are, not how they should be. It reveals facts, not norms. But one's knowledge that 1 + 1 = 2 has normative force. If two people enter what you thought was an empty house, and three people leave, you don't reconsider your belief that 1 + 1 = 2. You take that belief as fixed and, in light of it, reconsider your belief that the house was in fact empty. If it turned out that, according to a certain scale, rocks *R* and *R** had a collective weight of more than 2 pounds, even though you previously thought that each weighed no more than 1 pound, you'd take your belief that 1 + 1 = 2 as fixed, and would revise some other belief of yours (e.g., your views concerning the rocks' individual weights or the accuracy of the scale, or possibly even the laws of physics).

In fact, consequence of Relativity Theory is that if you add a rock weighing exactly 1 pound to a scale on which there is already a rock weighing exactly 1 pound, the collective weight of those two rocks will be *over* 2 pounds. Thus, weight is *not* a strictly additive property.[273] Neither, of course, is temperature. If you mix a quart of water with a temperature of 70° with a quart of water with a temperature of 90°, the result is not a body of water with a temperature of 160°. But we don't hold that 1+1>2 or that 70° +90°≠160°.

We interpret observations in light of the principle that 1 + 1 = 2, not *vice versa*, and the same holds for other 70+90=160 and, indeed, every other well-established precept of mathematics and, by obvious extensions of this reasoning, of logic.

3.4 Why (2) is wrong: The falsity of empiricism, a corollary of the falsity of conventionalism

First a terminological point. *P* and *Q* are ***consistent*** with each other if it's logically possible for both of them to be true; otherwise they're inconsistent with each other. So "grass is green" and "snow is white" are consistent with each other, as are "grass is black" and "snow is purple"; but "grass is green" and "grass has no color" are inconsistent with each other. A consistent set of propositions is one that doesn't contain any two propositions that are inconsistent with each other.

Conventions are indeed arbitrary. But whether the conventions one adopts are consistent with one another is not itself a matter of convention. It's a matter of convention that "1" denotes 1, that "+" denotes addition, that "snow" refers to snow, and so on. But, unless we changed those conventions, our linguistic conventions would form an inconsistent set if, without otherwise changing them, we added to them the convention that "1+1=3." For, our conventions being what they are, "1+1=3" entails that "1+1≠2" is true; and, our conventions being what they are, "1+1≠2" is false. So, even though it's up to us what our conventions are, it's not up to us, and therefore isn't a matter of convention, whether they're consistent with one another.

At the same time, we can't know on *empirical* grounds whether those conventions form a consistent set. To say that *P* is inconsistent with *Q* is to say that *P and not-Q* cannot *possibly* be true; and observations apprise one only of what is, not of what isn't, let alone of what cannot be. So, our conventions being what they are, it is neither a matter of convention nor an empirical fact that "1+1≠2" is false. It is a matter of non-empirical, non-conventional *logical* fact.

Even if we stipulate that *S* means *1+1=2* and that *S** means *1+1≠2*, it *still* isn't a matter of convention that those conventions are inconsistent with each other. There's nothing conventional about that fact that the number 2 cannot simultaneously have, and fail to have, the property of being the successor of 1. There is nothing human beings could possibly do to make 2 *not* be the successor of 1. We'll revisit this topic below.

The fact that consistency-relations are not a matter of convention means that there is nothing conventional about any body of statements that concern such relations. Mathematical principles are often, if not always, to the effect that some consistency-relation holds or fails to hold. Consider the proposition

that:

> (ET)[274] Euclidean triangles have interior angles whose sum is 180.

ET true. But there are no Euclidean triangles; the laws of physics prohibit it. So what is that proposition about? Non-existent triangles? No! Suppose that it *were* about such things. In that case, there would exist certain objects, viz. Euclidean triangles, that don't exist that it would be about. But that's absurd.

ET is to the effect that *if* an object is a Euclidean triangle, *then* its interior angles add up to 180. So ET is to the effect that, for any *x*, *x* is Euclidean triangle *is inconsistent* with *x*'s interior angles add up to 180.

Even *1+1=2* concerns consistency-relations, since, according to it, if *C* and *C** are non-overlapping one-membered classes, *then* the union of those classes is a two-membered class.

Conventionalism is rooted in a failure to distinguish between two very different statements, namely:

> (A) Among our conventions is one to the effect that sentence *S* means *P*

and

> (B) Our conventions being what they are, *S* means *P*.

Our conventions being what they are, it's settled what is meant by the sentence:

> (RC)[275] "if Richard Nixon were stronger than the Hulk, then he'd be stronger than Chuck Norris, unless of course Chuck Norris were much weaker than he is in actuality."

But there can't possibly be a convention concerning *that very sentence*. Before I wrote it down a few seconds ago, RC almost certainly wasn't even contemplated by anyone until just a moment ago, and even if a few people

did contemplate it, their doing so didn't give it the meaning that it has. This what *any* given complex expression means is partly non-conventional. In a moment, we'll show that this is indeed the case. Right now, we'll show that, *if* it's true, conventionalism is incoherent in a way that we have yet to point out.

"$(p/q)^2 = 2$" is a complex expression. So even though our linguistic conventions determine what is meant by "$(p/q)^2 = 2$", they do not do so directly. Supposing that P is the proposition affirmed by that sentence, it is a non-conventional *consequence* of our conventions that "$(p/q)^2 = 2$" has P for its meaning. Whether "$(p/q)^2 = 2$" is determined by what it means. (If it meant that snow is white, it would be true. But it doesn't, and it isn't.) So it isn't a matter of convention whether or not "$(p/q)^2 = 2$" is true. Since it therefore isn't a matter of convention what "$(p/q)^2 = 2$" means, it isn't matter of convention whether or not it's true. For whether a given expression is true is a function of what it means. "Snow is white" would be false if it meant that snow was black.

We will now see why anything whose meaning is directly fixed by convention is *ipso facto* a simple expression. If we *stipulate* that "$(p/q)^2 = 2$" has P for its meaning, then its having that meaning is no longer a function of what its parts mean. So, in that context, its parts are indistinguishable from things that don't have meanings and, in that context, are not functioning as expressions. Stipulating that a given expression has a given meaning strips its parts of meaning: those parts become meaningless homonyms of expressions with which, thanks to our stipulation, they have ceased to be identical.

It should be pointed out that whether a given sentence is true is only *partly* a matter of what it means, given that it's also partly a function of non-linguistic facts. ***Ceteris paribus***[***[276]***](#) "snow is white" would be false if snow were black, and *ceteris paribus* "snow is black" would be true if it meant that snow was white. *given* that snow is white.

4.0 Why *pre*perceptual and therefore non-empirical knowledge is a prerequisite for empirical knowledge

Innate cognitive structures are needed to have sense-perceptions and also to learn from them. Having a sense-perception means having a certain kind of mental representation of an object. Not all mental representations of objects are sense-perceptions. Mental images aren't, even though they resemble them. Concepts aren't, and they don't resemble them. It's a delicate matter exactly what kinds of representations are perceptual: a necessary condition, certainly, is that they be caused by their objects—your perception of Smith must be caused by Smith (by light bouncing off of him). It also seems necessary that (with certain qualifications) your perception "co-vary," or be capable of co-varying, in real time with the thing perceived. If you're looking at Smith, and he moves, you see him move (provided certain conditions are met; e.g., the movement isn't infinitesimally small). It's unclear whether there are other conditions that have to be met. (There probably are.)

In any case, perceptions are mental representations. And they represent their objects in four dimensions, and as having various chromatic (color-related), kinematic (movement-related), dynamic (force-related, causal) properties. Unless we take the implausible view that no intelligence is involved in the processes mediating between the somatic disturbances that lead to sense-perceptions, on the one hand, and those sense-perceptions themselves, on the other, none of this would be possible unless your mind, at some level, had some kind of pre-perceptual understanding of what it is for two things to be spatially, dynamically, etc., interrelated. So, in all likelihood, sense-perception *presupposes* knowledge.

4.1 Why non-empirical knowledge is a prerequisite for post-perceptual, inferential knowledge

We couldn't do very much with the information that sense-perception gave us unless we had non-empirical knowledge. First of all, your senses operate in the present. You are not seeing, touching, etc., whatever you're not *now* seeing, touching, etc. At most, you are remembering it. In order to apply what you remember to a current situation, you must see how what you remember *bears* on your current situation. You remember seeing somebody turn the key in their car and then driving away. You are now in front of a car with a key in it, and you want to drive it. Obviously you can apply what you remember to this situation. Let us suppose that you do so. So you think: "on *that* occasion, turning the key started the car; so the same thing will happen on *this* occasion. So, given that I want to start the car, I'll turn the key." Sense-perception (plus memory) *has* shown you that *on the first occasion* a certain car started after the key in it was turned. And, supposing that you turn the key in this car (the one in front of you), sense-perception will show you that another car started after the key in *it* was turned.

But sense-perception did not show you, and could not possibly have shown you, that there was a *relationship* between these two situations; it could not show you that, *because* the car in the first situation started after the key was turned, the same would be true of the car in the second. Your senses apprise you of situations; they thus give you *intra* situational information. They don't give you *inter* situational information. You can't see or otherwise sense-perceive relations between situations—between an event you saw a week ago and one you're seeing right now. Thus, any knowledge that you have of intersituational relations is partly (though obviously not entirely) non-empirical. Therefore any knowledge of *dependence*-relations holding between distinct situations is, at least in part, not strictly sense-based. Since it is only by virtue of knowing of such relations that one can know anything other than what one's senses are currently telling one, one can know extremely little if all knowledge is sense-based, and what one can know, one can know only for a fleeting instant.

4.2 Why non-empirical knowledge is a prerequisite for post-perceptual, inferential knowledge (continued)

Hume said that there is *no* legitimate inference from “that car (the one I saw two days ago) started when the key was turned” to “this car (the one I’m looking at now) will start if the key is turned” and, in general, that there is no legitimate inference from *any* proposition concerning the past to *any* proposition concerning the present or future. Hume acknowledges, of course, that past experience *makes* people form certain beliefs. His point is precisely that those beliefs, lacking as they do any legitimate basis, are not knowledge; they’re just compulsions, albeit ones that, as he grants, have thus far served us well for the most part. (See Chapter 11 for a discussion of Hume's views concerning induction.)

Although, unlike Hume, I believe that knowledge of the past *does* warrant beliefs about the present and future, there is an important truth in what Hume is saying. Hume has made it clear that *if* one permits oneself to use *only* such knowledge as one has acquired through the use of one’s senses, then one can’t infer anything about anything from anything. Inferences are about bearing relations. (More precisely, if one legitimately infers *Q* from *P*, then (i) *P* supports *Q*, (ii) one knows that *P* supports *Q*, and (iii) one’s acceptance of *Q*, given *P*, is driven by one’s knowledge that *P* supports *Q*.) If something about situation *X* is inferred from something about situation *Y*, that inference, if legitimate, must accord with some principle that is concerned with relations holding *between* distinct situations. And such relations obviously cannot, at least not as a rule, be sense-perceived.

To sum up, all knowledge that isn’t strictly sense-based—and this means all knowledge that involves any element of inference and, therefore, all theoretical knowledge (all theoretical knowledge is inferential, but not *vice versa*)—one must have knowledge of intersituational dependencies; and, since they can’t be sense-perceived, one’s knowledge of these dependences cannot be strictly sense-based.

5.0 The biggest problem with empiricism: thoughts not pictorial (even though pictures may facilitate thought)

What does perception give you? Images. (Not just visual images, of course, but also acoustical images, tactile images, etc.) If empiricism is right, this means that any thought that is not itself an image, or that contains information that cannot be encoded into an image, is unwarranted. But most thoughts, if not all of them, contain just such information. In other words, most thoughts, if not all of them, contain *non-iconic* information. For example, there is no image such that, in virtue of experiencing that image, one is thinking:

> (S) if Smith had fallen off the fence, he would have broken his leg.

What sort of image would represent *S*? Obviously you could have an image of Smith falling from a fence, and you can have an image of Smith breaking his leg. But *S* is to the effect that Smith's falling would *involve* his breaking his leg; it affirms the existence of a certain *dependence relationship*. And nothing could be an image of such a relationship. There can be images of rocks, trees, explosions, fallings, and waterfalls—but not of dependence relationships.

5.1 No iconic representation of dependence relations

Obviously there can be languages whose expressions consist of images. But so far as a given thing is functioning *as* an image, it isn't functioning *as* a linguistic expression. x is a linguistic representation of y only if x's representing y is a consequence of human conventions. Conventions can't make anything be an *image* of anything else. Conventions couldn't make a photograph of a waterfall be an *image* of a lion, even though such conventions could make such a photograph be a *representation* of a lion.

Expressions can be created by juxtaposing simple expressions. By juxtaposing "tall" and "man" one creates a third expression ("tall man"). But juxtaposing two images doesn't create a third image. Let X be what results when an image of Smith falling off a fence is placed to the right of an of Smith with a clearly broken leg. X is not itself an image. Therefore it isn't an image of the fact that that Smith would have broken his leg if he had fallen off a certain fence. We could stipulate that it represented that fact. But to the extent that X's representing that fact was a consequence of that stipulation, it wouldn't be an *image* of that fact.

True ***propositions*** are not responsible for their truth. (A proposition is any truth or falsehood.) If it's true, the proposition *Smith is in Paris* corresponds to some fact. If it isn't, it doesn't. We've seen that there cannot be images of the facts in virtue of which true conditional propositions hold. It follows that there cannot be images of disjunctive truths (those of the form *P or Q*). This is because, if true, a disjunctive propositions holds in virtue of the same fact as some conditional proposition. (*P or Q* holds in virtue of the same fact as *if not-P, then Q*. That is why the one cannot be true if the other is false.) It also follows that there cannot be images of the fact in virtue of which *P necessarily true* holds. (If true, *P is necessarily true* holds virtue of the same fact as *if Q entails not-P, then Q is false*, which is a conditional proposition.)

If Mary is tall, then she's not short says that one proposition (*Mary is tall*) cannot possibly be true if another proposition (*Mary is short*) is false; so it says that the first proposition's being true is dependent on the second one's being false. Thus, conditional truths are propositions *about* propositions. They are *meta-propositions.* (We'll refer to propositions that are not meta-propositions as "first-order propositions.")

Given *any* true meta-proposition, there can be no image of the fact in

virtue of which holds. *Snow is white* attributes a color to snow. But *it is likely that snow is white* doesn't attribute *any* property to *anything* spatiotemporal. It attributes a property (that of being likely to be true) to a proposition (that snow is white). P is likely to be true if, supposing that L is a list of the propositions we know to be true, the additions that must be made to L if P is true are less extravagant than those that must be made if P is false. So *P is likely to be true* says that one class of propositions (the set consisting of L plus those propositions we'd have to add to it were P true) bears a certain relation (in some respect having to do with the desiderata that an explanation must satisfy to be good) to another set of propositions (the set consisting of L plus those propositions we'd have to add to it were P false); and the first set bears that relation to the second if the first one is less different from L (in that same respect) than the second.

So the fact in virtue of which *P is likely to be true* holds involves the existence of instances of relations among relations among sets of propositions. Similar points hold fact in virtue of which any meta-proposition holds. So the reason such facts cannot be imaged is simply that nothing non-spatiotemporal can be imaged. Given that empiricism says that all knowledge is perception-based and therefore image-based, it's no wonder that empiricism has a hard time accommodating the fact that have knowledge of meta-truths.

5.2 Iconic information too specific in some respects and also too unspecific in others

First a terminological point: By a "general" thought, I mean one that concerns some unspecified thing (e.g., *somebody or other cured cancer*) or that concerns all things in a certain category without considering any one of them in particular (e.g., *most people have at least some decency*).

Suppose you believe that the cat is on the mat. Any *image* of a cat on the mat will include information not included in that thought (e.g., it will include information concerning the cat's shape or color, or concerning the mat's shape or size).

By the same token, any thought will contain information not included in any image. An image of a cat on the mat is indistinguishable from an image of a perfect cat-impostor that is on a mat. So if the thought that *the cat is on the mat* were identical with an image, it would be identical with the thought that *a perfect cat impostor is on the mat.* But those propositions are distinct.[277]

Also, general thoughts couldn't possibly be identical with images. Suppose you think that something or other is on the mat. An image can't just be an image of something or other on the mat; it must be an image of some particular thing on the mat. So if the thought that something were on the mat were identical with an image, it would be identical with the thought that (say) Snoopy was on the mat. But those thoughts aren't identical.

Similar arguments show that other general thoughts can't be identical with images. Consider the thought that *all* birds have feathers. An image of *all* birds —an image that depicted every single bird in existence—would be indistinguishable from an image of (say) 84 billion birds. But the thought *84 billion birds have feathers* is different from the thought that all birds have feathers. So neither thought is an image. Obvious adaptations of this argument show that no image can be identical with the thought that most birds/many birds/all possible birds have feathers.

So, really, *no* thought that can be expressed by a sentence is identical with an image. This shows that many thoughts aren't images and, since sense-perceptions are images, therefore have contents very different from those of any sense-perception.

5.2.1 Berkeley's unwitting anti-empiricism

Ironically, it was that greatest of empiricists, George Berkeley, who made it clear that general thoughts cannot possibly be identical with images. Berkeley's argument coincides with our own: any image is specific in ways that any one of those thoughts won't be. (Berkeley didn't make the correlative point that thoughts are specific in ways that images are not.) Berkeley saw that *if* all thoughts are images, then there are certain thoughts we can't have. Berkeley also saw that, if empiricism is correct, thoughts must be images. Finally, Berkeley saw that, given these two points, empiricism is inconsistent with the presumption that we can have general thoughts. But instead of rejecting empiricism because it has this absurd consequence, Berkeley held that, indeed, we can't have general thoughts. This position, in addition to being obviously wrong, is self-defeating, since the thought that there are no general thoughts is itself a general thought.

Berkeley qualified his position by saying that, although we couldn't *directly* grasp general propositions, we could grasp non-general propositions that we *took to stand for general ones.* But one can't take *x* to mean *y* unless one grasps *y*. So one can't take a non-general proposition to stand for a general proposition unless one grasps the latter. (See Sections 8–21 of Berkeley's "Principles of Human Knowledge.")

5.2.2 Berkeley's misuse of the term "idea"

Some philosophers believe that we are aware of universals and, therefore, that we are aware of non-spatiotemporal entities. Berkeley denies this. Berkeley says that any given thought is in its entirety concerned with "particulars" and that it is impossible to be aware of "universals."

A couple of terminological points before we proceed. Berkeley uses the term "particular" to refer to what, throughout this book, we have referred as "spatiotemporal" entities, and he uses the term "universal" to refer to what we have referred to as "properties" (or "universals" or "characteristics"). So "particulars," for Berkeley, are what we have referred to as *instances* of universals (or, to be precise, of *certain* universals (e.g., the property of being a dog), not all universals, since there are universals (e.g., the property of being an even number), whose instances are not spatiotemporal). And Berkeley uses the term "abstract idea" to refer to thoughts that (directly) concern universals or otherwise embody a (direct) awareness of them. Berkeley denies that we have abstract ideas.

Given his acceptance of empiricism, Berkeley is right to deny this, and he supports his position with arguments that, *relative* to the assumption that empiricism is correct, are maximally cogent. But not only do those arguments fail to establish that we don't have abstract ideas: they establish that we *do* have such ideas. We must make a preliminary point concerning Berkeley's use of the word "idea" before we see why the claims just made are true.

As he uses the word "idea," all "ideas" are thoughts and all thoughts are ideas. (In this context, I am using the word "thought" to refer to *any* case of anything's being aware of anything, it being irrelevant whether the mode of awareness is perceptual or conceptual and it being irrelevant whether the object of awareness is spatiotemporal or non-spatiotemporal.) But, for Berkeley, the sentence:

> (i) "all thoughts are mental images,"

expresses a substantive truth, not a linguistic convention. So, for Berkeley, (i) is in the same category as:

> (ii) "two is the unique number *n* such that *n* is an even prime iff it isn't possible for there to be a regular, 17-sided polygon,"

and it is *not* to be in the same category as:

> (iii) "if Smith has ten fingers, then Smith has 10 fingers."

Berkeley says that we cannot form "indeterminate" ideas. Here is his argument ("BA" is short for "Berkeley's argument[H11] ").

> (BA) All ideas (thoughts) are images. All images are "determinate." In other words, given any object and any characteristic, any image either ascribes that characteristic to that object or it fails to ascribe that characteristic to that object, and if the property in question is one that can be had to varying degrees (such as height), there is some specific degree to which the object in question is represented by that image as having the characteristic in question. So an image cannot just be of *a* man. It must be of a man who has a mustache; and an image cannot just be of a man who is rotund; it must be one of a man who is rotund *to some specific degree.* Thus, the idea that one could have an idea of "man" in general is preposterous. For such an idea would be an image of a man who had a beard *and* was clean shaven, who was six feet tall and five feet tall, who was standing *and* sitting, etc.

One problem with (BA) is that not all thoughts are images. In fact, discursive thoughts are *never* images. Imagery is often a *vehicle* for thought. When I think about philosophical problems, I tend to think about shapes and surfaces. But my having those images is neither necessary nor sufficient for my having those thoughts. A person's experiencing this or that series of images is never either necessary or sufficient for his having this or that thought. One oftentimes *uses* imagery to *direct* the course of a discursive thought-process, but imagery is never itself constitutive of such a process.

A consequence of empiricism is that a belief must be an image if is to be an instance of knowledge. According to empiricism, only beliefs that have a strictly observational basis are cases of knowledge. Observation gives us nothing images. So, if empiricism is right, information that cannot be embodied in an image has no place in any knowledge-conducive thought-

process. But *no* thought, and therefore no thought-process, consists solely of imagery.

Let *M* be some mental image that you are currently having of our friend Smith. *M*, let us suppose, is *not* a sense-perception. But for argument's sake, let's suppose that it is perception-*like* in that it is no less high-resolution or vivid than a sense-perception of Smith had under excellent conditions by somebody with excellent vision. In other words, *M* is about as "determinate" as a (non-perceptual) mental image can be.

Of course, if you *know* Smith's height, you are likely to *take* that image to one of somebody who is (let us suppose) six feet tall and relatively close by, and not as somebody who is six inches tall and very close by or 100 feet tall and far away. But it is only because you are reading background beliefs into it that *M* represents somebody who is six feet tall.

In fact, nothing in *M per se* could possibly specify Smith's height. If *M* represents Smith's height as coincident with that of some object that you know to be six feet tall, it is only because you are reading into *M* your background knowledge as to that other object's height. *M per se* leaves it open whether that object is six inches tall and very nearby or 100 feet tall and far away.

In fact, *M per se* doesn't represent Smith at all. Let *M** be a mental image that is qualitatively just like *M* but that occurs in the mind of somebody who doesn't know of Smith's existence and therefore isn't thinking about him. You are reading thoughts into *M* that this other person is not reading into *M**. But that is the only relevant difference between the two images. Considered strictly *as* images, *M* and *M** are indistinguishable. Obvious extensions of this reasoning show that, given *any* object, there is *no* image that *by itself* represents that object. Given some object *x* and some image *M#*, it is only if background information is read into *M#* that it represents *x*.

No image *by itself* represents anything. Images are often integral parts of representations. But it is only when supplemented by mental activity that doesn't consist of imagery that any image manages to represent anything. There is no object such that any image *per se* represents that object. *A fortiori* there is no object such that any image attributes *any* characteristics to it. It is the thoughts behind an image that determine what that image represents and, though possibly to a lesser extent, what it says about what it represents.

It is the thoughts behind an image that are determinate. Contrary to what Berkeley says, the image itself is completely indeterminate in respect of what

it represents and largely indeterminate in respect of what it says about it.

5.3 Knowledge is never wholly (or even partially) identical with mental imagery

As we observed, if all knowledge is strictly sense-based, then any case of knowledge is a case of some sequence of images running through one's mind. But *no sequence of mental images is knowledge.* Such a sequence may serve as a partial *basis* for knowledge (e.g., if the image-sequence is a series of visual perceptions of a dog, then, partly on the basis of that sequence, you may know that there is a dog near you); and one can have knowledge *about* that sequence (e.g., you can know that only a visually gifted person could experience such high-resolution mental images). Nevertheless, knowing something *always* involves grasping at least one thing in some way *other* than through an image and, therefore, in some way that, because sense-perception gives one nothing but images, empiricism cannot account for.

There's a difference between an image, on the one hand, and that image's being represented as *true,* on the other. (Philosophers use the word "veridical" to describe accurate perceptions, and they sometimes use the word "falsidical" to describe inaccurate ones. But I'll stick with "true" and "false.") Consider some painting (e.g., a painting of a woman picking flowers). There is nothing in the image itself that says: *this woman really exists, and she really is picking flowers.* Images are not self-referential; so they don't say "I am true" or "I am false." Thus, the image itself leaves it open to how it is to be taken. It could be taken as something true, something false, or something that is not meant to be taken as true *or* as false.

A corollary is that there is no image such that, merely by virtue of experiencing it, one is grasping the concept of truth. To know something, one must believe it to be true. Believing something to be true thus involves grasping the concept of truth and, given what we just said, therefore involves grasping something that cannot possibly be grasped by experiencing an image. Since the senses provide us only with images, it follows that the concept of truth is one that is not grasped through the senses. And since a grasp of that concept is a prerequisite to knowing anything, it follows that, contrary to what empiricism says, not all knowledge is strictly sense-based. In fact, it follows that, although a great deal of knowledge is obviously *largely* sense-based, *no* knowledge is *strictly* sense-based.

6.0 George Berkeley and the collapse of empiricism into idealism

George Berkeley (1685–1753) argued that "to be is to be perceived" (*esse percipi est*). He held, in other words, that the things we believe to be revealed to us *through* sense-perception (e.g., rocks, trees) *are* sense-perceptions—or, more exactly, are collections of sense-perceptions.[278] So Berkeley held that the book that you see before you is identical with the perceptions you (and others) have of it.

The view that physical objects *are* sense-perceptions is known as "idealism." It is so named because it is the view that external objects are "ideas," not because it embodies a spirit of optimism.

Even though, if our intuitions are to be given any credence, idealism is an abomination, Berkeley had a very good reason for accepting it. Given an acceptance of empiricism, one *must* be an idealist *if* one is to avoid saying that we have no knowledge of the external world at all. Berkeley saw this; and, being an empiricist who wasn't also a skeptic, he duly accepted idealism.

According to empiricism, if you don't know it through sense-perception, you don't know it. But what do you know through sense-perception? "I know that I'm reading a book right now." No you don't. You know that, *if what your senses are telling you is the truth*, you're reading a book.

Your senses cannot themselves give you any good reason to believe that they're truthful. Unless you already have some assurance that your senses are truthful, no attempt on their part to *tell* you this deserves any credence. So, supposing that you *do* know that you're reading a book right now, this knowledge, although *largely* based on sense-perception, isn't *entirely* so.

In general, even if our senses are generally telling us the truth, this fact cannot *itself* be known through sense-perception. This doesn't mean that it cannot be known. (I believe it can be known that our senses are telling us the truth. See Chapter 11.) But it means that, if we are to come to know that our senses are telling us the truth, it cannot be entirely on the basis of what our senses themselves tell us.

An analogy may help. Uncorroborated testimony is close to worthless. Smith is a suspect in a crime. He gives you an alibi. (You're the detective assigned to the case.) You want to know whether Smith's alibi is correct. If you ask Smith "are you telling me the truth?" he'll obviously say yes. What you want is *independent corroboration* for Smith's testimony. No person can self-corroborate; and neither can anything else, including sense-perception. If we are

to have any assurance that our senses are telling us the truth, that assurance must come from some non-sensory source. But empiricism says that nothing coming from any non-sensory source deserves any credit. Given this fact, there are only two (mutually exclusive) ways for empiricists, such as Berkeley, to go.

(i) Deny that we know anything about the external world.

(ii) Redefine the concept of an external object in such a way as to validate empiricism.

(i) is self-explanatory. It's skepticism. Some philosophers accept skepticism.[279] But for Berkeley, skepticism is not an alternative. He all but takes it for granted that it's false. (We'll say why in a moment.)

This puts Berkeley between a rock and a hard place. Berkeley refuses to stop being an empiricist. And he also refuses to accept skepticism.

This leaves him with only one option. He must show that the presumption that our senses give us knowledge of the external world can be accommodated *within* the limits set by the belief that all knowledge is strictly observational.

There is only one way to do this. One must reinterpret statements about the external objects as statements about one's own perceptions. In other words, one must take the view that the statement "that rock weighs five pounds" has the same content as some statement about one's own perceptions.

The position that statements about external objects can be translated into statements about perceptions is known as ***phenomenalism.*** Any attempt to translate a sentence about the external world into one about one's perceptions is known as a ***phenomenal reduction.***

Phenomenalism entails idealism. If every truth about rocks, trees, etc., is a truth about perceptions, then rocks, trees, etc., are nothing if not perceptions.

Phenomenalism is false. It simply isn't possible to translate statements about rocks (etc.) into statements about one's perceptions. We'll see why in Section 8.0.

But phenomenalism must be accepted by any true empiricist who isn't also a skeptic.

6.1 A systematic presentation of Berkeley's argument [280]

Step 1: If our senses are to be trusted, properties are observer dependent. A given elephant, at a given time, seems big to an ant and small to a giant. A given room, at a given time, seems cold to someone who just got out of a sauna and warm to someone who just came in from a blizzard.

Step 2: The external world is totally unknowable *unless* we suppose that, typically, things are as they appear.

Commentary on Step 2: This is a corollary of empiricism. Empiricism says: what we know, we know strictly on the basis of our senses. So if our senses tell us that *x* is big, then we're entitled to believe *x* is big—period.

One cannot, in general, legitimately second-guess what one's senses are telling one (at least not if one is an empiricist). One *can* throw out patently deviant perceptions (e.g., perceptions of suddenly appearing pink elephants). But this covers only an infinitesimal minority of our sensory experience.

Step 3: Things *are* knowable. It would be absurd to say that we don't know anything about rocks, trees, etc.

Step 4: Things can't have incompatible properties. A number can't be both even *and* odd. A thing can't be both big *and* small. A body of water can't be both hot *and* cold.

Step 5: Suppose, if only for argument's sake, that things exist independently of our perceptions of them. Suppose that, whether or not somebody perceives it, the vase exists (the same being true of the rock, the tree, etc.). In that case, things will have *incompatible properties.*

Explanation of Step 5: Remember Step 3—we *do* know about external objects. They are *not* unknown. So they *are* as they appear. Well, they *appear* to have incompatible properties. One and the same thing *appears* both big and small, hot and cold, fast and slow, etc.

Step 6: (this step is crucial to understanding Step 7, so bear with me): There's nothing incoherent in the idea that one and the same thing should

appear both big and small, hot and cold.

Clarification: Suppose I say “Bill is in Virginia right now” and I then say “Bill is not in Virginia right now.” Obviously there cannot be some one thing of which both statements are correct representations. But there’s nothing incoherent in the idea that both statements should be *made.* While it is incoherent to hold there is some one thing which they both correctly *describe,* it’s obviously perfectly ok to hold that both statements (however badly they may fit together) were affirmed. Similarly, there is nothing incoherent in the idea that, at the same time, there should occur a *perception* to the effect that *x* has one temperature and also a perception to the effect that *x* has some other temperature. What is incoherent is the idea that there should be some thing of which both perceptions are correct *representations.*

Step 7: If we say that our perceptions are *representations* of objects that lie on the other side of them, then (assuming, as we are, that our perceptions are accurate) things have incompatible properties. But if we say that our perceptions are *not* representations of such objects—if, instead, we say that perceptions *are* those objects—then things *don’t* have incompatible properties.

Clarification: Remember Step 6. It isn’t possible for there to be two correct but incompatible *representations* of some one thing, even though it’s obviously possible for two such representations to *exist.* Our perceptions give us incompatible messages (they say that one and the same thing is both big and small, is moving both quickly and slowly, etc.). So it isn’t possible for our perceptions to be correct *representations* of things. But it *is* possible for our perceptions to exist. It is irrelevant that they are incoherent if taken to be representations of trans-perceptual entities.

Step 8: It follows from Step 7 that our perceptions can’t be *representations* of anything.

Step 9: But remember Step 2—we *do* know about things. We *do* know their sizes, temperatures, etc. The way to make this all work is to say: things *are* our perceptions of them. We know our perceptions obviously. (So if we say that rocks, trees, etc., *are* our perceptions, then we can explain how we

can know them.) And we *also* know that perceptions can't be representations of things. Thus, perceptions don't *represent* reality: they *are* reality.

Conclusion: For something to exist *is* for it to be perceived. The perception doesn't represent the thing. The perception *constitutes* the thing.

7.0 Evaluating Berkeley's argument

Berkeley's argument fails. But the reason why it does must be understood in terms of a rather subtle point about sense-perception—a point that will initially be hard to understand but that, after being stated abstractly, will be made clear with some examples.

Our senses give us *relational* information. They tell us how things are in relation to one another and in relation to ourselves. They apprise us, not of absolute, but of relative heights, distances, velocities, temperatures, and masses; our eyes tell us, not that the car is going 35 mph, but that it is going faster than that other car; our muscles, when we exert ourselves, tell us not that the barbell weighs 215 pounds, but that it weighs more than the barbells we tried to move a moment ago. Our senses tell us, not that Smith is seven feet tall, but that he is much taller than Jones. And our senses tell us where things are, not in relation to some absolute coordinate system, but in relation to us. So the information given to us by our senses is relational in two senses: objects are related to one another; and they are also related to an egocentric, as opposed to universal, coordinate system.

Once this is taken into account, it is clear why Berkeley's argument crumbles. Consider a situation where *X* is walking at a rate of 1 mph, *Y* is driving at a rate of 30 mph, and *Z* is driving at a rate of 90 mph. (*X*, *Y*, and *Z* can see each other; they have good vision and normal cognitive capacities. So they know, at least approximately, what their relative velocities are.)

According to Berkeley, *X*'s perception of *Y* is simply this: *Y is moving fast*. And, according to Berkeley, Z's perception of *Y* is simply: *Y is moving slowly*. Berkeley rightly holds that both perceptions could be accurate. But he wrongly holds that the information embodied in the one perception contradicts that embodied in the other. This is because Berkeley has misrepresent both bodies of information. *X*'s eyes don't tell him: *Y is moving fast*. They tell him: *Y is moving faster than I am*. And *Z*'s eyes don't tell him: *Y is moving slowly*. They tell him: *Y is moving more slowly than I am*.

So when we consider what *X*'s eyes are telling him, and put it together with what *Z*'s are telling *him* (Z), the result is:

(i) *Z* is moving faster than *Y*, and *Y* is moving faster than *X*.

There is nothing incoherent about (i). It does indeed attribute two

properties to *Y* (namely, the property of moving slower than *Z* and the property of moving faster than *X*). But there is no reason why a thing can't simultaneously have both properties; for those properties are not incompatible—they aren't like the properties "square" and "circle." (i) describes a state of affairs that can, without any incoherence or self-contradiction, be attributed to the external world.

8.0 How we actually learn what properties things have

As we've noted, our senses give us only relational information; they tell us, not that the object *x* weighs 80 pounds, but that it weighs *more* than object *y*; not that object *x* is moving at a rate of 50 mph, but that it is moving fast*er* than *y*; not that *x* has a temperature of 78°, but that *x* is warm*er* than *y*; and so on. At the same time, we're obviously capable of learning objects' actual masses, temperatures, and such. But how is this possible if our senses give us only relational information?

To answer this question, we must take a brief detour through the philosophy of science. Suppose you decide to measure time in terms of your own heart rate. You'll find that, when you do vigorous exercise, everything in the universe (other than your heart rate itself) slows down. (Your heart rate does not itself slow down; for, being the standard of comparison, its rate is by definition constant.) And you'll find that, when you are calm and relaxed, everything in the universe speeds up.

Here you can say one of two things. (i) Because of some strange law, your exercising causes the entire universe (except for your heart itself) to slow down (and your ceasing to exercise causes it to speed up). (ii) You shouldn't measure rates of change with respect to your heart rate, since doing so creates otherwise absent causal anomalies and demands that relatively simple physical laws be replaced with exceedingly complicated ones.

Obviously (ii) is the right choice. You shouldn't regard your heart rate as a periodic process. You should say that it's your heart rate that's changing, not the entire universe.[281]

Comparable points hold with respect to all degree properties. (By a "degree property," I mean one that can be had to varying degrees. So coldness is a degree property, since one thing can be cold*er* than another. But the property of being divisible by two is not, since a number either has that property or it doesn't.) Suppose a consequence of taking object *L* to have a constant length is that, when *L* is put in a freezer, everything in the universe (except for *L* itself) expands and that, when *L* is put in boiling water, everything in the universe shrinks. In that case, obviously, you shouldn't take *L* to have an unchanging length; you shouldn't take it as your standard of comparison. The principle implicit in all this is that, if one learns that there has been a change in the difference between *x* and *y* in respect of how much

they have degree property phi, one should regard the extent to which *x* has phi as having changed *less* than the degree to which *y* has phi *if and only if* doing so is less costly, from an explanatory viewpoint, than it would be to regard *y* as having undergone less change (in that respect) than *x* or to regard *x* and *y* as having undergone equal amounts of change (same qualification). So, if fewer causal anomalies are created by regarding *y* as having undergone less phi-change than *x* than are created by regarding *x* as having undergone more phi-change than *y*, then one should *not* regard *x* as approximating, to a higher degree than *y*, to a condition of phi-uniformity. And if fewer causal anomalies are created by regarding *x* and *y* as having undergone equal amounts of phi-change than are created by supposing one of them to have undergone more phi-change than the other, then one should regard *neither* x *nor* y as approximating a condition of phi-uniformity to a higher degree than the other. *Otherwise* one should regard *x* as approximating to a condition of phi-uniformity to a higher degree than *x*.

8.1 How this last point relates to sense-perception

Our senses constantly give us information about comparative rates of change. One's eyes might tell one, for example, that the apparent size of the man is changing faster than the apparent size of the tree. The man's apparent size, so one's eyes tell one, is shrinking faster than the tree's apparent size, even though the apparent size of each one is shrinking. Without in any way being untrue to this raw, pre-inferential, visual information, one could hold any one of the following:

> (i) One's distance from the man remains constant, and so does one's distance from the tree. But the man and the tree are both shrinking, and the man is shrinking faster than the tree.
>
> (ii) One is moving *away* from both the man and the tree. (Therefore, one's distance from the man isn't constant, and neither is one's distance from the tree.) But the distance between the man and the tree remains constant; and, whereas the man is shrinking, the tree's size remains constant.
>
> (iii) One is moving away from both the man and the tree. But the distance between the man and the tree remains constant. And even though *both* the man and the tree are shrinking, the man is doing so at a faster rate than the tree.
>
> (iv) The tree's size remains constant, and so does the man's. The man is moving away from the tree in one direction, and you are moving away from the tree in the opposite direction.

Each of (i)–(iv) is consistent with the raw information contained in the previously described perception. But a world where any one of (i)–(iii) was correct would *ceteris paribus* be a world that was replete with causal anomalies absent from a world where (iv) was correct. This means that, if one were to countenance any one of (i)–(iii), one's understanding of the world and the forces that govern it would undergo a weakening that it would not undergo were one instead to countenance (iv). It is for this reason, then, that

ceteris paribus one countenances (iv), and not any one of (i)–(iii). It is for this reason, that, after exercising vigorously, one believes that one's heart rate sped up and not that the rest of the world slowed down; that, after lifting weights for a year, one became stronger and not that everything became lighter; that, on leaving a steam room, one's body has become warmer and not that the world that has become colder; and so on.

So although it is obviously *through* sense-perception that we learn objects' masses, speeds, temperatures, and so on, we don't *sense-perceive* them, at least not directly. Although it is through sense-perception that one learns that the water is 78°, one doesn't *sense-perceive* that the water has that temperature. What one sense-perceives is that the water is warm*er* than *x* and cold*er* than *y*.

Given this information, one is justified in believing that the water's temperature is 78° just in case, in doing so, one is undermining the integrity of one's understanding of physical reality less than one would be in not doing so. Basically, objects' actual properties are inferred from sense-perceptions and not directly sense-perceived; and the relevant rule of inference is: "of the various ways of accommodating the raw, sensory data that you've been given (and there will always be several such ways), the right one is the one that is the least *ad hoc* (that posits the fewest causal anomalies, that has the most explanatory fertility, etc.)."

9.0 Primary vs. secondary properties

Like John Locke, Berkeley distinguishes between ***primary properties*** and ***secondary properties.***[282] Examples of primary properties are shape, size, density, liquidity, solidity, gaseousness, and state of motion. Examples of secondary properties are sweetness, smell, color, and timbre.[283]

For a thing to have a given secondary property—e.g., for it to have the taste of Snicker's bar—is for that thing to be prone to induce sensations of a certain kind in creatures of a given kind.

For a thing to have a given *primary* property is for it to have an effect of a given kind on the conditions that objects must meet to occupy a certain region of space-time. By virtue of being round (or gaseous), a thing has certain effects on would-be occupants of its current location that, other things being equal, it wouldn't have were it square-shaped (or solid) instead.

According to some, the essence of Berkeley's argument for idealism is that primary properties are really secondary properties.[284] Thus interpreted, Berkeley's argument is as follows:

> Berkeley's argument (BA): For a thing to have a given secondary property is for it to taste, smell, feel, etc., a certain way. Some philosophers hold that, in this respect, shape, size, and other so-called primary properties are fundamentally different from secondary properties. But they're wrong. Shape, size, etc., are quite as observer-dependent as taste, color, etc. When you're near the house, it looks big (we're assuming that your perception is accurate); and when you're far from the house, it looks small (same qualification). Of course, we *could* say that the size an object *seems* to have is different from the size it *actually* has. But in that case, we'd be stuck saying that our senses tell us nothing, or next to nothing, about the external world, which is absurd. It follows (for the reasons given earlier) that the objects of sense-perception are identical with our perceptions of them.

The problem with BA is that, contrary to what it says, your first house-perception *is* consistent with your second house-perception. The first perception doesn't say that the house is *big.* It tells you how its size compares

with those of various objects. It tells you that it's shorter than the telephone pole next to it, but taller than the cactus in front of it, etc. The comparative information given to you by your second perception is consistent with it. The relative apparent sizes of the cactus, house, and telephone pole are the same.

In the second perception, the house appears the way that small things appear *when looked at from up close.* And *that* is the one and only sense in which the house "looks small." But you're not looking at the house from up close. You're looking at it from far away. And your perception isn't telling you that the house is close by and looks that way. It's telling you that you're far away from the house and it looks that way. So it isn't telling you that the house is small. It's actually telling you that it's quite big.

When you're near the house, your house-image hogs up your entire field of vision. It occupies more space on your visual map. When you're far from the house, your field of vision has to include a great many objects *other* than the house. So it doesn't eat up very much map-space. Thus, measured in terms of how much map-space they eat up, the first house-*image* is larger than the second house- *image.*

But that doesn't mean that those perceptions make different statements concerning the size of the house. On a map the size of a postage stamp, the discoloration that represents Denmark will be miniscule. On a map the size of the side of a big house, the corresponding discoloration will be much bigger. But that doesn't mean that Denmark is small according to the small map and big according to the big one. Each of those maps tells you that Denmark's actual size corresponds to the *relative,* not the *absolute,* size of the picture of Denmark on that map. The bigger that discoloration is *in relation* to the other nation-representations on that map, the bigger the size thereby attributed to Denmark. Thus, the size attributed to Denmark by the small map might be twice that attributed to it by the big map.

The same thing *mutatis mutandis* is true of visual images. A big house no more "looks small" when looked at from far away than Denmark "looks small" on the tiny map. When the house is looked at from far away, the corresponding house-*image* is small (i.e., it eats up a comparatively small amount of visual space). But that's consistent with that very perception's *representing* that very house as being huge.

9.1 BA embodies a spurious understanding of what secondary properties are

The health effects of eating a foul smelling cut of meat are very different from the health-effects of eating a wholesome smelling cut of meat that is otherwise observationally just like the first. The way a piece of meat smells is an indication of its ability to nourish. For this reason, otherwise hard to obtain causal knowledge is easily obtained and otherwise hard to make predictions are easily made.

Differences in secondary properties *always* correlate with causal differences. In many cases, day-to-day experience makes it clear what causal properties are associated with possession of a given secondary property. (Day-to-day experience makes it clear that things with certain tastes and smells are not fit to drink.) In other cases, day-to-day experience doesn't do this. (Day-to-day experience doesn't make it clear how an object that is red differs in respect of its causal properties from one that is green but is otherwise identical.) But if a given object has a given secondary property, it *ipso facto* has some causal property that it wouldn't otherwise have.

For this reason, secondary properties are not in any relevant sense "subjective." The term "subjective," as John Searle (1992) points out, has two very different meanings. To say that such and such is "subjective" can mean that:

> (i) its existence presupposes that of some subject,

or it can mean that:

> (ii) it embodies prejudices or other subject-specific idiosyncrasies that tend to inhibit the acquisition of knowledge.

My belief that 2 + 2 = 4 is subjective in sense (i). That belief presupposes the existence of a subject. There can't be disembodied beliefs floating about. [285] But that belief isn't subjective in sense (ii). It is (so we will assume) based on rational consideration of the relevant information. By contrast, my intense fear of people who wear plaid shirts (which is based, not on rational consideration of the relevant data, but on childhood traumas that we needn't discuss) is subjective in sense (ii). In other words, it is rooted in peculiarities

of my own person that make it harder for me than it would otherwise be to see things as they are. My olfactory perception that the maggot-infested meat in front of me smells awful is subjective in sense (i) but not in sense (ii). It is subjective only in the trivial and irrelevant sense in which *any* mental entity, even a coolly made judgment, is subjective: it is subjective in the same way as my rational and informed belief that 2 + 2 = 4 is subjective. But it isn't subjective in the same way as my prejudice-driven fear of people who wear plaid shirts.

Bearing this in mind, suppose that, all of a sudden, the way things smelled was no indication as to whether they were rotten or not. So, for example, a putrid-smelling cut of meat might or might not be good to eat—its smell would give you no information as to its fitness as a source of nourishment—the same being true of a wholesome-smelling cut of meat. In general, we wouldn't know anything about an object's causal properties from the way it smelled. But we often know a lot about a thing's causal properties from the way it smells. Smells correlate *extremely* reliably with certain causal features of objects—sometimes much more reliably than other properties. Sometimes the way a thing smells is the only indication (short of eating it or sending it to a toxicology lab) that it's laced with poison or that it's completely rotten.

Smells track primary properties; they track facts about microstructure that would otherwise be hard to detect. If a police-detective suspects that *X* is pure cocaine, he can verify his hypothesis by (i) snorting *X*, which would have various adverse consequences, (ii) sending *X* to a forensics laboratory, which would be extremely slow, or (iii) tasting it, which would be quick and easy.

A relevant fact about method (ii) is that many scientific techniques would have been impossible to develop had there not been reliable correlations between microstructure and odor. This is true of secondary properties generally.

There are some *apparent* counterexamples to our contention that secondary properties track microstructure. For example, a thing's color doesn't correlate reliably with its microstructure. From far away, the ocean water looks black; from up close it's transparent. Locke concluded from this that secondary properties are subjective.

But this is not the only conclusion to draw. An object's apparent color is a function *both* of that object's microstructure *and* of the relevant observer's physical relation to that object. But so does an object's apparent shape, apparent size, smell, etc. So the variability of apparent color no more entails the subjectivity of color than does the variability of apparent shape. And so long as

the conditions of observation are held constant, an object's apparent color does reliably correlate with its microstructure, just like its apparent shape (though not, it must be admitted, to as high a degree).

The right conclusion to draw, I believe, is that what we refer to as an object's color is an objective property, not of the object *per se*, but of the object's *relation* to environment in which a given observer is perceiving it. Thus, color is neither subjective; nor is it an objective property of objects *per se*; it is an objective property of ordered triples. If *O* is some object, *P* is someone perceiving that object, and *R* is *P*'s physical relation to *O*, then there is a property, to which we refer as "the color of *O*," that, though not actually a property of *O* itself, is an objective feature of the world, being property of the ordered triple <*O*, *P*, *R*>.

Setting aside people with visual defects, there is little or no variation between people in regards to how one's physical relationship to a given body of water affects how that water appears. Given two people who are free of visual pathologies, if that object appears bright blue to the one person when he bears relation *R* to it, that object will appear bright blue when the other person bears *R* to it. Of course, the color that a given object *appears* to have to a given person is a function of changes in that person's physical relation to that object that haven't the slightest effect on that object itself. But an object's *apparent* color does *not* vary, so long as the observer's physical relationship to it is fixed. A given object's physical constitution doesn't determine its apparent color. But *relative* to an observer bearing a fixed relation to that object, that object's apparent color *is* fixed. This suggests that what we refer to as an object's color is a property of *relations* that observers may bear with respect to that object. Thus, colors are properties of ordered triples of the earlier-described kind.

If this is right, then instances of color are mind-dependent but opinion-*independent.* In other words, there would be no instances of colors if there were no minds. But it isn't a matter of opinion what color such an instance has. There are many facts that are mind-dependent but opinion-independent. Right now, I am having various thoughts while standing within 20 feet from where Smith is standing. It isn't a 'matter of opinion' where I am at a given time in relation to Smith or what I'm thinking at that time.[286] But (supposing that I couldn't exist without a mind) neither fact is mind-independent.

9.2 BA embodies a spurious understanding of what secondary properties are (continued)

Our *perceptions* of primary properties are very different from our perceptions of secondary properties. When one sees a porcupine, one's visual image has a structural similarity to that of the porcupine; given any two features of the porcupine that are distinguished *by* image, there are corresponding differentiations *in* that image.[287] Perceptions of secondary properties are very different; they *aren't* internally differentiated, at least not in a way that at all corresponds to the objective facts to which they are hewed.

This, presumably, is at least part of the reason why it's widely held that an object's secondary properties are really properties of the subject's experiences of that object, and not of that object itself. (It's also why it seems more natural to speak of "sensations" of taste, smell, etc. than of "perceptions" of them.) But, I will now argue, an object's secondary properties belong to it no less than its primary properties. I must warn that the forthcoming argument is quite intricate and that, for the first few paragraphs, it will seem as though I'm arguing for the contention that an object's secondary properties *don't* really belong to it. I must ask for a certain amount of patience on the reader's part.

Let *S* be the smell that, in actuality, is had by extremely putrid meat; and let *S** be the extremely pleasant fragrance that, in actuality, is given off by eucalyptus trees. There is no *inherent* reason why eucalyptus trees couldn't have S and there is no inherent reason why rotten meat couldn't have *S**. In other words, if there occurred a smell-exchange between eucalyptus trees and rotten meat, we would not, at least not for that reason alone, forfeit the causal knowledge that we now have by virtue of knowing that a cut of meat has *S* or that a eucalyptus tree has *S**.

Of course, there would *initially* be a loss of knowledge. One would mistakenly think that the pleasant-smelling cut of meat one was about to eat was not rotten. But given some time to adjust to the new smell-situation, one would cease to make such mistakes.[288]

9.3 BA embodies a spurious understanding of what secondary properties are (continued)

There is another reason why, even though it wouldn't have any lasting effect on our ability to acquire knowledge through olfaction, that smell-exchange would initially compromise our ability to do so. This fact *initially* suggests that olfaction doesn't give us knowledge of the causal structure of the world. But, duly scrutinized, it turns out to entail that olfaction is, like vision and audition, *does* give us such knowledge.

Much of the knowledge that we acquire through the sense of smell is based on the comparative properties of our olfactory sensations. Other things being equal, the more rotten a cut of meat is, the worse it smells. Other things being equal, the more pungent the fragrance of a eucalyptus tree is the richer that tree is in the chemical responsible for that smell.

Bearing this in mind, suppose that *S* is the exact smell had (under a given set of circumstances) by a cut of meat that was rotten to degree *n*, and that *S** is the fragrance (same qualification) of a eucalyptus having amount m of the relevant chemical. Because the previously described exchange hasn't occurred, given a cut of meat having *S* and some other, otherwise observationally identical cut of meat, with a much stronger (weaker) version of *S*, I know on that basis that *ceteris paribus* the latter cut of meat is more (less) rotten than the former. For this reason, if the exchange in question occurred *but that other, compensatory changes didn't also occur alongside*, our ability to acquire comparative knowledge of this kind would be diminished.

But there's no reason why any given set of causal properties should be associated with any given type of olfactory sensation. What *is* necessary is that, once the assignments are made, their mutual relations parallel the mutual relations of their objects.

> (SA) So you admit that its having this as opposed to that odor does not, in and of itself, diminish our ability to learn about the world through the sense of smell. You admit, in other words, that rotten meat doesn't *have* to have the smell it does. But in admitting this, you're admitting that it doesn't *really* have that smell.

There are two problems with SA.

Problem #1: When we say that a given cut of meat has a bad smell, *we're talking about the cut of meat.* If the word "smell" referred to *sensations*, then sentences like "that meat has an awful smell" wouldn't be about the meat. By the same token, so far as that sentence *is* about the meat, it isn't about sensations. Thus, to the extent that it really is about the meat, whether or not that sentence makes a true statement has nothing to do with anyone's sensations. And to the extent that it isn't about our sensations, "that meat has an awful smell" is about the meat's causal properties or, in any case, is about some other (probably related) mind-independent fact.

Problem #2: SA establishes that smell is mind-dependent *only* to the extent that it establishes the obviously, or at least presumably, false conclusion that size and shape are mind-dependent.

Look at the book in front of you. Let *V* be the visual sensation you are having. Under otherwise unchanged circumstances, you could have a visual sensation that, although very different from *V*, gave you either the very same information or that gave you much more information. In human beings, the set of sight isn't nearly as acute as it is where certain other species are concerned. Were they to look at the book in front of you, under conditions just like the ones that actually obtain, they'd have sensations very different from *V* that contained all the information contained in *V*, and then some.[289] You could have visual sensations very different from the ones you actually have without, for that reason, failing to know anything that you do know.

But this doesn't entail that the facts of which vision apprises us are mind-dependent. *V* gives you a lot of information about the book's location, its shape, and its size. And *V*'s shape, size, etc., are quite definitely mind- *in*dependent.

This reasoning is likely to meet with the following objection:

(QB)[290] But aren't you begging the question against Berkeley in saying this? Wasn't Berkeley arguing that, since (as you just pointed out) secondary properties are relevantly similar primary properties, to be is to be perceived?

Berkeley does indeed argue that. But he also takes it for granted that secondary properties are mind-dependent. In other words, his argument has the form: (i) secondary properties are mind-dependent; (ii) secondary properties are relevantly similar to primary properties; (iii) therefore, primary properties are mind-dependent; (iv) it follows from (iii) that objects are mind-dependent—are creatures of the mind.

But I argued that secondary properties *aren't* mind-dependent. I reject step (i), and provided an independently plausible reason for doing so. Therefore, what I've said doesn't beg any questions against Berkeley and QB thus fails.

10.0 Is temperature a primary or a secondary property?

Some properties that seem to be secondary turn out to be primary. A brief examination of one such property will clarify some of the points just made.

For an object to have a certain *temperature* is for the particles composing it to have a certain mean kinetic energy. (Basically, it's for those particles to move around in a certain way. The faster the movements, the greater the temperature. [291]) But it's initially tempting to categorize temperature with sweetness and, thus, to see it as a secondary property.

There are a couple of reasons for this. First, secondary properties are modality-specific. Sweetness can be tasted, but not heard, seen, touched. By contrast, primary properties are clearly transmodal. Squareness can be seen or touched. In fact, creatures with very good hearing (e.g,. bats), can *hear* shapes, just as we can see them; bats can detect an object's shape through sonar. And creatures with a very olfactory keen sense (e.g., moles) can *smell* shape.

According to some, including Locke and Berkeley himself, temperature can be known only through touch. But this isn't true, of course. Given what temperature is (namely, molecular motion), it is just as capable of being known through sight or audition as it is of being known through touch. But it's obvious why someone would think otherwise.

There's another reason why temperature is often falsely regarded as a secondary property. There isn't anything inherently pleasant or unpleasant about seeing a square-shaped object or a fast-moving object. Under certain circumstances, it may be very unpleasant to see something move with a certain speed. But the connection is circumstance-specific. There is, however, something inherently unpleasant about certain odors, tastes, and timbres. In general, primary properties are not inherently associated with aesthetic properties—any such association is circumstantial—whereas secondary properties *are* non-circumstantially associated with such properties. Very high and very low temperatures are non-circumstantially associated with unpleasant sensations, and moderate temperatures are non-circumstantially associated with pleasant ones.

Nonetheless, temperature is a primary property, since an instance of temperature is identical with various instances of motion. And, technically, temperature *is* transmodal; for one can learn a thing's temperature by figuring out how its parts are moving, which obviously need not, and often cannot, be established on the basis of the "feeling of temperature" on the basis of which

one determines that one's beverage is warm or that it's too cold to go for a run.

11.0 A second Berkeleyan argument—"to be is to be conceived"

At one point, Berkeley argues that to be is to be conceived ("esse concipi ist"). In other words, things are the awarenesses that we, or other thinking beings, have of them. Here is his argument:

> (BA2) You cannot conceive of an unconceived object without its *ipso facto* being conceived. So it is literally inconceivable that there should be unconceived objects. (You can't conceive of an unconceived chicken, since, by virtue of your conceiving of it, it isn't unconceived.) Thus, the very idea of an unconceived object makes no sense, and unconceived objects no more exist than four sided triangles.

If cogent, BA2 establishes that being conceived of is a *necessary* condition for existing, but not that it's a *sufficient* condition for it. (Given only that one can't conceive of a circle that isn't a closed figure, it doesn't follow that all closed figures are circles.)

In any case, BA2 spurious. The problem with BA2 that, contrary to what Berkeley assumes, *propositions*, not *objects*, are the objects of conception. When one "conceives of a 50-pound chicken," one is thinking about a proposition (namely, *there exists a chicken that weighs 50 pounds*). There isn't some specific chicken *x* such that *x* weighs 50 pounds and such that one is thinking about *x*. Similarly to conceive of an unconceived chicken isn't for there to be some chicken *x* such that one is conceiving of *x* and such that one is conceiving of *x* as being unconceived. Rather, one is thinking about the proposition: *there is a chicken x such that nobody is conceiving of it*, which one can grasp, and even know to be true, without there being any chicken *x* such that one is conceiving of *x*. So BA2 fails to show that *to be is to be conceived.*

12.0 The phenomenal reduction—some preliminaries

First some terminological points. An ***object-statement*** is one that concerns mind-independent entities; e.g., "that statue is over 10 feet tall" is an object statement. A ***perception-statement*** is one that concerns one's own perceptions (e.g., "I am having an experience that, if it were veridical, would be of a statue that is over 10 feet tall"). Two statements are ***observationally equivalent*** if there is no observation that (dis)confirms the one either more or less than it (dis)confirms the other. (All analytically equivalent sentences are observationally equivalent—but not vice versa, as we're about to see.) Finally, a property or relation is ***observationally inert*** if no instance of it could possibly affect what anyone observes.

12.1 The phenomenal reduction

For the reasons given earlier, strict empiricism leads to either skepticism or to idealism. Skepticism and idealism are both deeply at odds with commonsense. Berkeley agreed that skepticism was at odds with commonsense. But he denied that idealism had this defect. He held, in fact, that idealism *is* the commonsense view.

In an effort to vindicate this obviously false claim, Berkeley argued that any object-statement that we could have any good reason to believe correct can be reinterpreted as a perception-statement.

Here is a modernized presentation of Berkeley's reasoning:

> (BR)[292] A statement about the external world is meaningful only to the extent that, were it true, we'd have observations that we wouldn't otherwise have.
>
> This entails that such a statement is meaningful only *to the extent* that it concerns observations. When the possible observations run out, so does the meaning. Thus, "Smith has a car" is without meaning to the extent Smith's having a car is observationally inert, the same thing *mutatis mutandis* obviously being true of "Brown has a car." And:
>
> (1) "Smith's car weighs twice as much as Brown's car,"
>
> is meaningful only to the extent that, for any objects *x* and *y*, if *x* weighs twice as much as *y*, there would *ipso facto* occur observations that would not otherwise occur.
>
> If we replace all of the constants in (1) with variables, what results is the sentence-form:
>
> (2) *A* bears *R* to *B*.
>
> A true sentence results if the variables in (2) are replaced with constants referring to classes of perceptions. Let K_1 be the class of perceptions that one has in virtue of the fact that Smith's car exists. (K_1 might include a perception of Smith

pulling out of his garage.) Let K_2 be the class of perceptions that one has when, and only when, there are objects *x* and *y* such that *x* weighs twice as much as *y*. And let K_3 be the class of perceptions that one has in virtue of the fact that Brown's car exists.

If a given observation *O* confirms (1) to a degree, it confirms:

(3) K_1 bears K_2 to K_3,

to that same degree. (In this context, "confirms" is to be taken to refer to positive or negative confirmation.) If a given observation *O** (dis)confirms any one of these three claims, it is *ipso facto* (not) a member of one of K_1–K_3. So there can't be an observation that (dis)confirms (1) more or less than it (dis)confirms (3).

There is therefore no observation-based reason to prefer (1) to (3). So assuming, as we are, that empiricism is correct, and that all good reasons are observation-based, it follows that there is no reason at all.

But *the very fact that there is no such reason* is itself a reason to prefer (3) to (1). (3) doesn't posit anything for which there is no observational evidence. (We'll henceforth use the word "trans-observational" to describe entities for whose existence there is no observational evidence.) But (1), at least as many philosophers interpret it, *does* posit trans-observational entities.[293] And to the extent that a sentence demands the existence of trans-observational entities, it is meaningless. Why? Because sentences are meaningful only to the extent that, if true, they make a difference; and, given the truth of empiricism, a sentence makes a difference only if it makes an *observational* difference.

12.2 The viability and significance of the phenomenal reduction

In this context, I'll use the word "perception" to denote any mental state that, *if veridical*, is an accurate sensory observation but that needn't be veridical. Thus, hallucinations are perceptions, as we'll be using that word, and so are veridical perceptions.

Note: in what we're about to say, words like "see" and "perceive" are not to be taken as "success-verbs." As I'll be using the word "see," the fact that you "see" Smith driving a car doesn't mean that you *really* saw him driving a car. It means only that you had the sort of experience that *would* have been veridical had Smith *actually* been driving his car and had that in fact been what you were seeing. Given any other similar term----"perceive," "perception," "hear," etc.---the same point *mutatis mutandis* holds with respect to it. Having cleared that up, let's consider the following hypothetical situation.

You see Smith driving a car. Berkeley admits this. In other words, Berkeley admits—as he surely must—that, other things being equal, your having a perception of Smith driving a car gives some weight to the contention that Smith is driving a car.

Question: *How* is *P* evidence of Smith's driving a car? It's evidence of it because it *says* that Smith is driving a car. If your mother, who you trust, calls you up and says "Smith is driving a car (which he shouldn't being doing, since he's blind drunk)," you're likely to believe that Smith is driving a car, even if you didn't previously have any reason to believe it. The reason is that you regard your mother's testimony as evidence to the effect. You also regard your visual perception of Smith driving as (extremely strong) evidence to that effect. You do this because your eyes are telling you—much as your mother might, but much more convincingly—that Smith (or, at any rate, somebody who looks a lot like him) is driving a car.

Part of what your eyes (or, if you're hallucinating, the malfunctioning brain-centers) are telling you is that some spatially remote object (namely, Smith) bears a certain relation (namely, that of driving) with respect to some other spatially remote object (namely, the car Smith is driving). And in telling you that Smith and his car are spatially remote from you, your eyes are saying, or at least implying, that Smith isn't a part of you or, therefore, of your mind.

It may be that *P* is wrong. It may be that you're hallucinating and that,

when you come down from your acid trip, it will turn out that Smith wasn't driving at the time in question. But *P* still *said* that he was driving, and it still *said* that there exists some spatially remote object bearing a certain relation to some other such object.

Any given perception is to the effect that objects that aren't a part of you bear certain relations to one another and/or individually have certain properties. So to the extent that our sense-perceptions are evidence of anything, they are evidence of trans-perceptual objects. And that is why statements about perceptions cannot possibly be equivalent with statements about objects.

13.0 Two insights of Berkeley's

Berkeley's argument for idealism does not go through, and neither does his attempt to reduce statements about objects to statements about perceptions. But, in the process of trying to establish idealism, Berkeley had two extremely profound and original insights, whose philosophical consequences are still being fathomed, to wit:

> (A) Many of the "statements" we make aren't really statements at all, and are instead *statement-forms,*

and

> (B) A statement's grammatical form may not coincide with its logical form—that, in other words, a statement must be reparsed if its meaning is to be made clear.[294]

(A) *explained:* "The number five is odd" is a statement. It attributes a definite property to a definite object. The expression "*x* is odd" is a statement-form, not a statement. An "interpretation" of a statement-form is simply an assignment of constants to the variables. To ***model*** a statement-form is to identify an interpretation of it that yields a true statement. Here is a trivial example:

> (1) A bears relationship *R* to *B.*
> (2) If *x* bears *R* to *y*, *y* bears *R* to *x*.
> (3) Nothing bears *R* to itself. In other words, ‹*x* bears *R* to *x*› is false, for all objects *x*.

Let *R* be the relationship that one bears to somebody else in virtue of being that person's next-door neighbor. Since (setting aside the odd person who owns adjacent houses) nobody is one's own next-door neighbor, (3) comes out true. So does (2), since *x* must be *y*'s next-door neighbor if *y* is *x*'s next-door neighbor. Thus, supposing for argument's sake that Smith and Jones are next-door neighbors, (1) comes out true if "*A*" refers to Smith and "*B*" refers to Jones. We have thus modeled (1)–(3).

In conjecturing that (so-called) statements such as 'that rock weighs five

pounds' can be reinterpreted as statements about perceptions, Berkeley was in effect saying that such statements aren't statements at all—they're statement forms. Though he didn't himself put it this way, he was saying (i) that the terms in them (e.g., 'rock,' 'weighs five pounds') are really variables, and (ii) that true statements would result if those variables were replaced with constants referring to perceptions.

Berkeley was wrong about (ii), as we've already observed. But he *was* right about (i). There is no one microstructure that Rover must have if my current perception of him is to be accurate. There is, as we will now see, no one way anything must be if *any* given perception is to be correct.

We've seen that our knowledge of the world is strictly comparative.[295] We know how fast a given thing is traveling *only* to the extent that we know how much fast*er* it is traveling than some other thing. We know how heavy/hot/big/etc., a thing is *only* to the extent that we know how much heavier/hotter/bigger/etc., that thing is than some other thing. Spatiotemporal knowledge is comparative knowledge.

That means that *any*thing—no matter what it's parts are 'made of'—whose parts are appropriately interrelated is no more, and no less, consistent with what our sense-perceptions tell us than is any other thing whose parts are thus interrelated. This means that perception apprises us only of *structure.*

Our perceptions (if correct) tell us the structure of the world, and our physical theories (if correct) apprise us of micro*structure.* Given *any* two things having the requisite structure, there cannot possibly be any observational basis for holding that it, and not the other thing, is what is really out there.

This means that, if *x* and *y* both have the requisite structure but are qualitatively different in some non-structural respect, the supposition that *x* is what is really out there models the data exactly as well—no better, no worse—than the supposition that *y* is what is really out there.

This doesn't mean that there is no fact as to what is out there. There obviously is. Given any set of possible objects, each having the structure of the external world but each differing in some non-structural respect from the others, at most one of those objects can be the external world. But the supposition that any given one of them is what is actual does exactly as good a job of modeling the observational data as the supposition that any other given one of them is what is actual. And one of the claims embodied in Berkeley's argument(s) for idealism is precisely that, to the extent that the probability of a hypothesis is a function of its congruence with observational data, anything that models the

observational data *ipso facto* is no less likely to be what's really out there than anything else that does so. Berkeley was therefore (so far as I know) the first model-theorist.

(B) explained: To my knowledge, Berkeley was the first to recognize the distinction grammatical structure (surface structure) and logical form (deep structure[296]). To avoid getting caught up in technicalities, let's also suppose Smith is the only percipient being in existence. Berkeley's idealism seems to entail that, if Smith closes his eyes (and is otherwise not in sensory contact with the world), the whole world disappears; and it also seems to entail that the whole world springs back into existence when Smith opens his eyes. This is deeply implausible.

It also seems to undermine Berkeley's claim that any object-statement that is presumed correct (e.g., 'the world doesn't disappear when I close my eyes') is equivalent with a statement stating the observational evidence for that claim. When I close my eyes, the *observational* evidence seems to suggest that the world has vanished. And it's only, so it is thought (or so I personally think), because one knows that one's observations *aren't* the one and only relevant yardstick that one knows that the world hasn't disappeared.

Berkeley's response to this is brilliant. He says that the real meaning of: [H12]

> (1) "The world doesn't go out of existence when Smith (who, for argument's sake, we'll assume to be the only sentient being in existence) closes his eyes,"

is ***perspicuously***[297] represented by a sentence whose surface-structure bears no resemblance to (1)'s. That other sentence is one along the lines of:

> (2) "If Smith closes his eyes for *x* minutes and then reopens them, the things he sees are where one would *expect* them to be, *given* that they are subject to the usual forces."

Berkeley's own illustration of this principle is particularly effective. The stars are moving, Berkeley points out. A more precise way of putting it, as Berkeley himself makes clear, is that we are moving relative to the stars, and the stars are therefore moving relative to us. But, as Berkeley points out, we don't *see* the stars move. The blips in the sky that we see on a given night

aren't in the same relative positions as the blips in the sky that we see on other nights.

For this reason, we do a kind of "regression analysis" on the blips that we see over the years. We try to connect the blips seen on night 1 with those seen on night 2, and we try to connect the latter with those seen on night 3, and so on. So if $N_1...N_n$ are nights in question, and n is some high number (e.g., 500), we identify curves such that, if it is assumed that those curves are the trajectories of those blips, it follows that those blips obey well-defined, relative simple laws. So the logical form of:

> (A) "The stars are moving even when nobody is looking at them,"

is perspicuously given by the grammatically very different sentence:

> (B) "If it is assumed that, when we're not looking at them, the stars obey certain laws and therefore have certain trajectories, we can make observationally accurate predictions that we couldn't otherwise make, and arrive at observationally correct explanations that we couldn't otherwise arrive at."

Berkeley said, as clearly as it can be said, that what sentences really mean isn't what, given their overt forms, they seem to mean. So setting aside some petty and insubstantial objections relating to the exact meanings of "grammatical" and "logical," Berkeley was, to my knowledge, the first person to recognize the distinction between grammatical and logical form.

Part IV Determinism, Causality, and Probability

Chapter 13 Determinism, Randomness, and Unpredictability

1.0 Determinism and indeterminism

Determinism is the view that how the world is at any given point in time fixes how it will be in every respect at all later times. So given how the world is now, there is only one way it could possibly be in a thousand (or a billion) years. And given how the world was a minute ago (or a billion years ago), the world *has* to be the way it is now: things couldn't possibly be any other way.

Indeterminism is the doctrine that determinism is false. So it is the view that, for at least *one* moment in time, how the world is at that time does not completely fix how the world will be at all later times.

Indeterminism is *not* the view that the world is chaotic or lawless. Indeterminism is actually an extremely moderate doctrine. All it says is that, at *some* point in time (it could be past, present, or future), some event occurs that doesn't have to occur. So, technically speaking, for indeterminism to be true, there needs to be only one "renegade" event.

Determinism is thus much harder to establish than indeterminism. After all, the determinist says that, for *any* time t, how the world is at t determines how the world will be in *every* respect at *all* later times. Which is a very strong claim. Indeterminism, on the other hand, only says that there is *at least one* counterexample to what the determinist says. Which is not a very strong claim.

Nonetheless, even though determinism is a strong statement for which it is difficult to produce definitive grounds, we know that it is true or at least a very close approximation to the truth. It is an open question whether, at the sub-atomic level, the world is strictly deterministic. But whatever indeterminacies there are at that level seldom, if ever, percolate up to the molecular, cellular, biological, planetary, or astronomical scales. So, for most intents and purposes the world in which *we* operate—the world of "medium sized" objects—is deterministic.

To be sure, the world often *seems* to be indeterministic. But careful scrutiny of any such situation, with the possible exception of those at the sub-atomic level, tends to show that it is governed by deterministic mechanisms.

2.0 Determinism vs. predictability

Predictability and determinism are not the same thing. If a system is completely *in*deterministic, then it is indeed unpredictable (or only fortuitously predictable), since prediction requires regularity and there is regularity only where there are deterministic laws. But a deterministic system may be unpredictable. In fact, even if the world is completely deterministic—even if the state of the universe at any point completely fixes how everything will be, in every respect, at all later points—there are limits to how much any creature could predict about the course of future events. It doesn't matter how intelligent that creature is or how sharp its senses are or how good its technology is. For reasons of logic, no creature is able to predict everything that happens. There are four reasons for this.

2.1 Why predictability ≠ determinism: Reason #1[298]

We learn about the external world through *sense-perception*—that is, through vision, hearing, touch, etc. Let's focus on vision for the moment. Suppose that you're seeing some rock. (Let's refer to that rock as "*R*.") Here's what's involved in your seeing *R*. Some light beam bounced off *R*. (We'll refer to that light beam as "*L*.") As a result of that collision, *L*'s structure changed in such a way that *R*'s structure is now coded in its (*L*'s) structure. For this reason, *L* bears information about *R*. *L* hits your eyes, disturbing them. Those disturbances reflect *L*'s structure. Given that *L* reflects *R*'s structure, it follows that these disturbances of your eyes reflect *R*'s structure. Those ocular disturbances therefore encode information about *R*. The resulting neural impulses transmit this information to some part of you that decodes this information. The results of this decoding are embodied in a visual perception that succeeds in representing *R*.

Light is energy. Things change when bombarded by energy. They don't always change very much. But there is always *some* change. It takes time for light, or any other information-bearing process, to travel. During that time-interval, the changes in *R* that were *initiated* by *L* led to other changes in *R*. Since, for the time being, *L* is your only source of information about *R*, and since *L* gives you dated and otherwise inaccurate information about *R*, the information encoded in the visual perception that *L* caused you to have is itself inaccurate.

Before we take the next step, I'd like to clarify this point. Let *t* be the time that you have the visual perception of *R* that *L* caused you to have. Obviously *t* will happen after the time *L* hits your eyes and *t* is later than the time that *L* collided with *R*. During the time between that collision, on the one hand, and *t*, on the other, *R* has changed. Why has it changed? Because the just-mentioned collision made it change. That collision, as we just pointed out, set off processes because of which *R* changed.

What you know, at *t*, about *R* is confined to what *L* tells you. And what *L* tells you is dated information—it's yesterday's news. What you know, by virtue of having the sense-perception in question, is now dead and gone. What it tells you is how *R* was *a little while ago;* it doesn't tell you how *R* is now.

Supposing that *R* is some rock in your backyard—and not some rock in the depths of outer space—the time lag is minimal. And, no matter where *R* is, *L* did extremely little to change it. The changes that *L* induced in *R*, by colliding with it, fall well below the limits of sensory discrimination.

But mentally replace *R* with something small—something so small that the photons involved in its being bombarded with light *do* appreciably change it. In *that* sort of case, the thing you see at *t* is *very* different from the thing that is now out there in the world.

You might be aware of this gulf between the perception and the reality. But any attempt on your part to narrow this gulf will only make it wider. If you try to learn more about *R*—by, for example, shining more light on it—you'll make it *more* unlike the thing it was initially. And you'll only compound the problem if you try to undo those distortions with counter-distortions. Setting aside the problem just discussed, there's no point in trying to reverse the effects that *L* had on *R* by shining more light on *R*, since you don't know how *R* was before you bombarded it with light. Therefore, you don't know how *R* has changed and, consequently, you don't know what the needed corrections are.

What is true of your seeing *R* is true, to some degree or other, of *anything's having any sense-perception of anything.* It doesn't matter whether the percipient beings in question are super-brilliant aliens or iguanas or God Himself. As a matter of logic, perceiving a thing involves your being on the receiving end of causal process (e.g. a light-beam) that collided with that object and, as a result, was so changed that its structure came to reflect that of the object. Of course, that collision changed didn't just change that

causal process: it also changed the object. But you can't possibly know what those changes were, since they occurred *after* the relevant information-transmitting process left that thing. (Let's assume that process to be a light ray.) For the time being, that light ray is your one source of information about that thing. So whatever that light ray did to that object—whatever changes, if any, it made—you can't know what they are. If no changes occurred, you can't know it; and if changes did occur, you can't know that either, and a *fortiori* you can't know the exact nature of those changes. So what you see is different from what is out there, in the relevant place, at the time of your perception. And, as we saw a moment ago, attempts to remove this margin of error by having more perceptions will only make it bigger.

We are not talking about the limitations of human beings. We're taking about the conditions that must, as a matter of logic, be fulfilled by any sense-perception on the part of anyone or anything. If it's said that God could "just know" what is out there—that He didn't have to be on the receiving end of some causal process starting with external objects—I would reply by saying: "rubbish!" If there's no causal process, then, by definition, there's no information for God to go off of. And if there's no such information, then any image God comes up with is a hallucination. It might be a correct one. But, if so, it will only be an *accidentally* correct one, and it won't be enough for knowledge of anything, and any belief He has in consequence of having that image will be nothing more than a guess. A correct guess, perhaps. Some guesses are lucky guesses. But lucky guesses aren't knowledge. (See Chapter 10 for an explanation of this.)

If it's said that the laws of logic don't apply to God, I would say: "Yes they do." If, in response to that, it's said that God, being omnipotent, can't be hemmed in by anything, even the laws of logic, I would reply by saying that this statement embodies a total failure to understand the nature of logical truth—that God's inability to create a square circle is no more an abridgment of His powers than an athlete's "inability" to score a touchdown in a soccer match shows a lack of athletic talent. The idea that God could *know* of some event without having any causal connection is in the same category as the idea that a sufficiently good athlete could score a touchdown in the course of a soccer match. It's simply incoherent.

But why exactly is it incoherent? Only *justified* true beliefs are knowledge. Lucky guesses don't count. A justified belief is one for which there is evidence. One spatiotemporal thing is evidence of another if and only if the

first is causally related to the second in such a way that, given the existence of the first, that of the second can be inferred. Therefore, one's own mental states are evidence of a given external thing, and therefore give one a justification for believing that thing to exist, *only* if those mental states are effects of that thing. So the idea that God could know of some spatiotemporal entity without being causally connected to that entity is no more coherent than the abject absurdity that one event could be evidence of another even though the occurrence of the first event provided no reason to suspect the occurrence of the second. Since the latter idea is obviously a non-starter, which deserves to be put in the same category as "squares don't have four sides," the same is true of the former.[299] For similar reasons, one would simply be redefining the terms "'justification" and "knowledge" were one to say that unjustified beliefs were knowledge or that beliefs for which one had known evidence were justified. Nothing that follows from definitional truths can be false; anything that follows from a truth of logic is itself a truth of logic. So it is a truth of logic, since it follows from truths of logic, that God couldn't have knowledge of some spatiotemporal entity without being causally connected to that entity. It's also a tautology that causal connections are disruptive and that tautology entails that God's mental representations of the world must, in at least some cases, be discrepant with the corresponding realities.

To sum up: Learning about the world involves changing it. A consequence is that the world is different from how it is represented to you as being. So even if the world is 100% deterministic, there is necessarily a gulf between the perception and the reality.

Conclusion: Given the existence of the aforementioned gulf, any predictions one makes are necessarily slightly off. A prediction involves two things: first, knowledge of initial conditions (how things are at the moment—or, in any case, how they were just a moment ago); second, what the relevant causal laws are. Given knowledge of those causal laws, one can predict what must occur, given those initial conditions. But since one's knowledge of initial conditions is necessarily off, though sometimes not by very much, one's predictions will be off.

2.2 Why predictability ≠ determinism: Reason #2

What we know affects how the world is. Why is this? Because what we know affects what we do, and what we do affects how the world is.

We can't know what it is that we'll know. Suppose that, right now, you know exactly what it is that you'll know in 10 years. (Let *K* be that body of knowledge.) In that case, what you'll know in 10 years will be *K plus* all the knowledge you pick up between now and then. And if you build *that* knowledge, viz., the knowledge you pick up along the way, into your definition of *K*, then you'll *already* know it and you won't pick it up along the way anymore, and the knowledge that you *do* pick up along the way will be something else. So it makes no sense to say that you could know now what it is that you'll know in 10 years.[300]

At the end of this chapter, we'll discuss some other reasons why a perfectly deterministic world is necessarily *not* one in which, for any time *t*, any creature, no matter what intellectual or sensory virtue it had, could possibly predict everything that would happen after *t*.

2.2 Why predictability ≠ determinism: Reason #2 (revisited and developed)

Given a complex object, one part of it can be aware of another part; but, at any given time, no one part can be aware of itself. For a person to be "self-aware" is for one part of his mind to be aware of another part of some other part of his mind. During a reflective moment, there must be a part of your mind that is not an object of reflection. It may be that you can be aware of a part of your mind that, in its turn, is aware of some third part of your mind. It may be that instances of "second-order" self-consciousness are possible. It may even be that, for any number n, self-consciousness of order n is possible (logically, if not psychologically). But, at any given time, there must be some part of one's mind that one is not aware of. If, at a given time, some person were aware of *everything* in his mind, then he'd be aware of his awareness of everything in his mind, and also of his awareness of that awareness, and also his awareness of *that* awareness, and so on *ad infinitum.* Such a mind is not only a psychological, but also a logical, impossibility.

The part of one's mind that one is not aware of is no less a determinant of future events than any other part of one's mind. This means that there is necessarily a part of one's mind, and thus of some determinant of future events, of which it is impossible that one should be aware. An obvious corollary is that one's inability to be perfectly self-aware limits one's ability to predict the future.

2.2.1 Why predictability ≠ determinism: Reason #2 (continued)

A necessary condition for complete knowledge of the future is complete knowledge of the present. But if one had perfect knowledge of the present, one couldn't make predictions; indeed, one couldn't do anything. It is inherent in the concept of intentional conduct that the agent be ignorant of both some aspects of the future and some aspects of himself.

If I know that, at time *t*, forces over which I have no control will force me against my will to go to the store, then I cannot possibly decide that, at *t*, I'll make a trip to the store. If I know that if I'm at home on Tuesday forces over which I have no control will force me against my will to go to the store, I can decide to go to the store on Monday, so as to spare myself the trauma of a compulsory store-trip on Tuesday. But if I know that, at a given time, forces over which I have no control will transport to me to the store then I cannot possibly *choose* to initiate a journey to the store that begins that very time. One cannot possibly *choose* to make an event occur (at a given time) if one knows that (at that time) that event will occur anyway.

At the same time, one cannot choose to try to prevent an event whose occurrence one knows to be inevitable. One can try to prevent an event whose occurrence one believes to be *almost* inevitable. I can *try* to make a hole in one as long as I think there is *some* chance of my succeeding. But if I know that there is simply no chance of my making a hole in one, I cannot, as a matter of logic, even try to make one.

Thus, one can make choices only to the extent that one doesn't know what will happen. This *by itself* doesn't show that we can't have complete knowledge of the future. For it could be maintained that, when one makes a choice, one eliminates the very doubts concerning the future that one had to have in order to make that choice. (I'm in doubt as to whether I'll go to the store before, but not after, I choose to do so.)

But this move is to no avail. There are two reasons for this: a trivial one and a non-trivial one. The trivial one is that, in making a given decision, one seldom, if ever, eliminates all of the uncertainties that made that choice possible. We don't typically know that we'll manage to carry out the decisions we make, and we have little knowledge as to how, whether successfully carried out or not, those decisions will affect the world.

The non-trivial reason is that one cannot possibly know what choices one will make. This is a corollary of the fact that one cannot try to do what one

knows to be inevitable or to prevent what one knows to be impossible. “But surely that’s false,” it will be said. “I can know that, at some point tomorrow, I’ll choose to smoke a cigarette.” In describing oneself as knowing that one will choose to smoke a cigarette tomorrow, one is saying, in a distorted way, that one has *already* chosen to smoke a cigarette tomorrow. To choose whether or not to smoke a cigarette is to make up one’s mind one way or the other. Knowing that tomorrow one will definitely make up one’s mind to smoke a cigarette is the same as having already made up one’s mind to smoke a cigarette tomorrow. So, although one can predict that one will smoke a cigarette tomorrow, that is only because one has *already* chosen to do so; it isn’t because one has foreknowledge of some choice not yet made. Choices are *ipso facto* incapable of being predicted (by those who make them) and so, therefore, are their effects (to those who make them).

2.2.2 Why predictability ≠ determinism: Reason #2 (continued)

The word "predict" is ambiguous. Sometimes it means *have a belief about the future* and sometimes it means state a belief about the future. The fact that creatures have beliefs about the future entails that a perfectly deterministic world is not completely predictable, and so does the fact that creatures sometimes *state* their beliefs about the future. But statements about the future impose even greater limits on what can be known about the future than do unstated beliefs about the future, and the limits imposed by such statements are, up to a point, qualitatively different from those imposed by unstated beliefs.

Two stories will illustrate these points.

> *First Story*: *X* is a company that specializes in medical research. It employs a team of brilliant scientists who will make many important discoveries *if*, and *only if*, *X* stays in existence. Billionaire Smith knows that he'll live another 20 years or so if company *X* stays in existence and that he'll soon die if it doesn't. Smith also knows that, unless he manipulates the market in highly illegal ways, *X* will instantly collapse. Without telling anyone of any of this, Smith manipulates the market in the needed ways, knowing that, as a result, he'll live another 20 years and also that otherwise never-to-be-discovered cures will be discovered. Smith cannot possibly know the exact nature of what he'll learn in those 20 years, and *X*'s research-team cannot possibly know what *it* will come to know.

> *Second story*: Smith correctly believes that he'll live another 20 years if *X* stays in existence and that he'll soon die if it doesn't. Smith correctly believes that *X* will make a speedy economic recovery *if*, and *only if*, he publicly announces that it will do so. (Such an announcement, as Smith knows, will stimulate investment in *X*, thus keeping it afloat.) Knowing that he'll make the announcement, Smith knows that *X* will prosper and that, as a result, he'll live another 20 years. Smith cannot

> know what he'll come to know in the time that, by making the announcement, he will buy for himself, and *X*'s research-team (once it finds out that it won't perish) cannot know what *it* will come to know in the time that Smith has bought for it.

In the first story, knowledge of future events led to unforeseeable events. In the second story, somebody's *stating* that he had such knowledge led to such events. *Statements* about the future make the future unpredictable in one way. *Beliefs* about the future make it unpredictable in a different way. Since the word "prediction" sometimes means "belief about the future" and sometimes means "*statement* about the future," the thesis that the future is "unpredictable" is ambiguous; and even though both disambiguations are true, they're true for very different reasons.

2.3 Why predictability ≠ determinism: Reason #3

In this section, we will set aside the fact that, because of interference effects, there are limits to the degree of precision with which we can know what happens at the sub-atomic level. We will assume, falsely and strictly for argument's sake, that there *is* a way to eliminate the gap between:

(i) How the world is,

and:

(ii) How the world is represented to us that is the inevitable result of the fact that knowledge is acquired by interacting with the world.

We will see that, *even if this assumption is granted*, a strictly deterministic world is *still* one in which there are limits to what any being---no matter how intelligent, no matter how unencumbered by fatigue, lack of acumen, limitations of memory and life-span, etc.---can know about the future. More specifically, we'll see that, for reasons inherent in the very concept of *instantaneousness*---i.e., of what an instant in time is----it cannot be known exactly what is happening at *any* given instant in time.

If we're right about this, it obviously follows that we cannot know exactly what is happening at the present instant. Given that any knowledge we have of the future is based on our knowledge of current conditions, it follows that, for reasons having nothing to do with interference effects, there are limits, even in a strictly deterministic universe, to what it is theoretically possible to know about the future.

2.3.1 Why predictability ≠ determinism: Reason #3 (continued)

What does it mean to say that the car's speed *at instant t* is 60 mph? Speed = distance/time. An instant is an infinitely short stretch of time.[301] The time-interval encompassed by an instant is nil. So the car cannot therefore cover *any* distance during that interval. It follows, so it would seem, that it makes no sense at all to speak of the car's speed *at* an instant.

To say that the car's speed *at t* is 60 mph is to say that the car's speed during arbitrarily short stretches of non-null, *t*-inclusive time-intervals approximates to 60 mph to an arbitrarily high degree. If the car's average speed during *t*-inclusive interval S_1 is 62 mph, there is some shorter *t*-inclusive interval S_2 such that the car's average speed during S_2 differs from 60 mph by *less than* 2 mph. And if S_3 is some *t*-inclusive time-interval during which the car's average speed is 60.0001 mph, there is some shorter *t*-inclusive interval S_4 such that the car's average speed during S_4 differs from 60 mph by less than 1/10,000 mph. And so on.

All statements about what obtains during a given "instant" are elliptical for statements about what obtains during non-null stretches of time. If *O*'s mass is constantly increasing, to say that, at instant *t*, *O*'s mass is *M* is to say that, given any non-null *t*-inclusive time-interval S_1, if *O*'s mass deviates from *M* by a given amount, there is a shorter *t*-inclusive interval S_2 such that *O*'s mass deviates from *M* by less than that amount.

Measuring an object's speed involves seeing how much distance it covers during some interval of time. There is no other way---or, more precisely, no other *direct* way—of determining a given object's speed. Given some object *O* whose speed at time *t* one wishes to ascertain, one cannot directly ascertain *O*'s speed *at t*. (One can *in*directly ascertain it, as we'll see shortly.) What one *can* directly ascertain is *O*'s *average* speed during some *t*-inclusive interval. The smaller the interval, the better an approximation to *O*'s speed *at t*. But---and this cannot be stressed enough---it is only by considering *O*'s speed during (*t*-inclusive) *stretches* of time that one can determine *O*'s speed.

2.3.2 Why predictability ≠ determinism: Reason #3 (continued)

That said, there may be reason to believe that *O*'s average speed during increasingly small *t*-inclusive intervals approximates without limit to 60 mph. (Put more precisely: there may be reasons to believe that, given any number *x*, no matter how small, there is some number *y* such that, if one considers a *t*-inclusive time-interval of length *y*, the difference between 60 mph and *O*'s velocity at *t* is less than *x*.) Supposing for argument's sake that one establishes that *the smaller the t-inclusive interval that one considers the closer to zero is the difference between O's speed during that interval and 60 mph, then one has thereby established that O's speed at t is 60 mph*. Indeed, the italicized claim *is* the meaning of "*O*'s speed at *t* is 60 mph." (See Section 2.5.1.) It's not that, because of our scientific limitations, establishing the truth of the italicized claim is the best we can do in the way of establishing that *O*'s speed at *t* is 60 mph. It's that, for reasons inherent in the concept of what an instant is the two claims are identical, the former being but an abbreviation for the latter.

Since any direct (non-inferential) measurement of an object's velocity will be of an object's velocity during some *stretch* of time, any judgments one makes as to that object's speed during some instant will involve, as we'll now see, a heavy element of inference. Suppose that, during non-null, *t*-inclusive interval INT_1. Bearing this supposition in mind, consider the following two statements:

> (i) *O*'s average velocity is 60 mph.
>
> (ii) if INT_1, INT_2….are *t*-inclusive intervals such that, for any *n*, INT_n is bigger than INT_{n+1}, *O*'s average velocity during INT_{n+1} is closer to 60 mph than it is during INT_n.

One is justified in accepting (i) if, and only if, one is justified in accepting (ii).

But how could (ii) possibly be established? Not directly, obviously. Given any time-interval, one can measure *O*'s average velocity during that interval but once. Speed isn't like length. One can measure an object's length over and over again, since length is a property of enduring objects. But one can't

do this with speed, since speed is a property of non-recurring events. So *if* one judges that *O*'s speed at instant *t* is 60 mph, that is based *partly* on the fact that during the *t*-inclusive interval during which one measured *O*'s *average* speed, that (average) speed was close to 60mph.

Assuming that judgment to be made on good grounds, there is another, rather troublesome dimension to it. For, in making that judgment, one assumes that:

> (DI)[302] had one done the equivalent of performing the doubly impossible task of considering an infinitely long series of *t*-inclusive intervals, each smaller than its predecessor, one would have found that the limit of the series to be 60 mph.

Why, incidentally, would the task mentioned in DI be "doubly impossible"? Because:

> (a) since an event only occurs once, one can measure an object's speed no more than once,

and:

> (b) for a variety of reasons, it's impossible to make *infinitely* many measurements.

2.3.3 Why predictability ≠ determinism: Reason #3---taking stock before closing the argument

Before closing the argument, let us take stock.

(i) one can directly ascertain *average* speeds, but one cannot directly ascertain instantaneous speeds.

(ii) Fact (i) has nothing to do with our limitations as scientists; it has to do with the very meaning of statements of the form 'such and such was the case *at* instant *t*.' (Why this is so will be clarified and defended in the next section.)

(iii) For any time *t* and any object *O*, one cannot really *know* what *O*'s speed *at t* is; one can, at most, make reasonable but highly theoretical, and therefore inherently tentative, inferences as to what *O*'s speed at *t* is.

(iv) The inferences mentioned in (iii) themselves involve assumptions as to the instantaneous speeds of objects and, unless those assumptions are constituents of a theory for which there is considerable independent evidence, are viciously circular.

2.3.4 Why predictability ≠ determinism: Reason #3---Closing the argument

It is inherently impossible to *know* with what speed an object is traveling *at a given instant.* ("But doesn't the speedometer on my car indicate how fast my car is traveling *at a given instant*?" The speedometer on your car *does* (consistently, non-accidentally) tell you how fast your car is traveling at arbitrary instant *t if* various theories turn out to be right; e.g., theories correlating the car's speed at a given instance with the rate at which the engine turns over. Each of those theories is based on many a belief to the effect that, for some object *x*, velocity, and instant *t*, *x is moving at v at instant t.* Those beliefs don't register data; they embody judgments that are no less theory-dependent and no more secure than your judgment that *O*'s velocity at *t* would coincide with its average velocity during the *t*-inclusive interval that you considered.

In conclusion, setting aside anything having to do with interference effects and also with human intellectual and technological fallibility, *it is impossible, for reasons inherent in the concept of instantaneousness, to know the state of the universe at any given instant.* This means that we cannot have maximally precise knowledge of the present. And, given that we know both the past and the future on the basis of *inter alia* the present, this in its turn means that we cannot have maximally precise knowledge of the past or the future.

What we just said about an object's velocity at a given instant is true of its mass, length, color, and all of its other explanatorily significant properties. We cannot have direct knowledge of an object's mass at an instant; at most, we can make reasonable inferences as to its mass at a given instance on the basis of our knowledge of its average mass during some interval that includes that instant. To be sure, there is an important difference between speed, on the one hand, and mass and length (etc.), on the other. Speed is a property of non-repeatable events involving objects. Mass and length (etc.) are properties of objects *per se.* But this makes no difference in the present context. Given some general truths concerning the nature of measurement, which we're about to state, along with the points already made, it will be clear why this is so.

3.0 Determinism vs. randomness

"Random" doesn't mean the same thing as "undetermined." A deterministic world isn't necessarily one in which nothing random happens; and, give or take a few niceties, an indeterministic world isn't necessarily one in which anything random does happen. A story will make it clear why these points hold.

Smith arrogantly insulted a high-ranking member of a particularly lethal criminal organization, and that person has thus sworn to kill him. For unrelated reasons, Smith's wife also wants him dead, and she has just hired a hit man to kill him. She has always wanted him dead; she just married him for his money. Plus, Smith has a serious heart condition, and his doctor is surprised that he didn't die years ago. You are a friend of Smith's, and you know all of this. But you don't know anything else that is relevant to an estimate of how much longer he is likely to live. Given that information, the right estimate is, of course, "not long."

But Smith beats the odds. Nobody kills him and nothing goes wrong with his heart. But one day, as he's making a sales pitch to some prospective clients, a ceiling-lamp comes loose and falls on his head, killing him.

This was a "random" occurrence. But it wasn't *uncaused.* When a ceiling lamp falls, there's a reason for it (e.g., the bolt wasn't screwed in tightly enough or the weight of the lamp was too much for the beam it was hanging from). It was "random" because, although caused, it wasn't the sort of thing that, given the information at your disposal, could reasonably be expected to do him in. Thus, it was "random" not because it was *un*caused, but because it was *independently* caused, meaning that the causal sequence that led to his death was independent of the causal sequences that were under way that *were* reasonably believed, given the information available to you, to eventuate in his demise. So, in this context, "random" doesn't mean "uncaused," it means "uncaused by anything that we had reason to believe would occur."

This is what "random" typically means. It may be that microphysicists deal with uncaused events. But the rest of us deal exclusively with things that are quite definitely caused. (As previously stated, the indeterminacies at the micro-level don't percolate up to the macrolevel, at least not enough to engage the structures that change human lives.) So the "random" things we deal with are things that have unforeseeable causes.

A consequence is that randomness isn't a property of events; it's a relation

that an event bears to a body of knowledge. Relative to what you know, it is unforeseeable, and therefore random, that a ceiling lamp will kill Smith. But relative to what Howard the building inspector knows, that same event is not random. (Howard knows that the plank in question will go at any time, and he sees Smith standing underneath it.)

Thus, randomness isn't a property of events. It's a property of *statements*—or, more accurately, of relations between statements. An event is random, relative to a certain body of information, if its occurrence couldn't reasonably be predicted on the basis of that information. So randomness is really a relation that holds between the statements describing some body of data and the statement describing some event; that relation holds if the former statements don't warrant acceptance of the latter.

It's obvious that there can be random events in a deterministic world. Smith's death was an inevitable consequence of pre-existing circumstances. But, relative to the information that you had, it was also quite random.

3.1 Determinism vs. randomness (continued)

Supposing that the world is *indeterministic,* would there for that reason be "random" events in it? The answer is "yes." But the reasons are not what one would think; and the amount of randomness in such a world is not necessarily as great as one would expect it to be.

For argument's sake, suppose we know that, given the current state of the universe, it's causally completely undetermined whether, in five seconds, object *O* will move to the right or to the left. Each possible outcome is as likely as the other (they're "equiprobable").There's no way, of course, to predict which way it will go. We could make that prediction only to the extent that we had reason to think one of the outcomes the more likely one—which we obviously don't. But when, at the moment of truth, it goes to the left, was that "random"? We couldn't predict it; but it isn't exactly random. On the contrary, it's consistent with our expectations of it.

To see why, let's modify our earlier story about Smith. This time, he isn't killed by a ceiling lamp. The vindictive criminal kills him. Before Smith died, you knew that, in all likelihood, he would be killed *either* by that criminal *or* by his wife *or* by his heart-condition—but you didn't know which, since, given what you knew, they were all equiprobable. When the criminal gets him, that surely isn't random. Quite the opposite. Even though it could not itself be predicted with certainty to happen, it belonged to a class of events which, given the available evidence, was known to contain all and only the things that *would* kill him.

The same is true of *O*. It is known that, in 5 seconds, O will either to go the left or to the right. So even though its going to the left as opposed to the right could not be predicted, its going to the left wasn't exactly random. So indeterminism doesn't entail randomness.

But here we must make a subtle distinction. *Relative* to the knowledge that it's undetermined whether *O* goes to the left as opposed to the right (or *vice versa),* its going to the left is random. So *O*'s going to the left is both random (relative to one body of knowledge) and non-random (relative to another).

4.0 Why determinism, if true, is capable of being verified

On hearing me say that the world is at least approximately deterministic, students of mine often respond by saying either that determinism is unverifiable or that it's simply false. Depending how it's meant, this statement is either false or vacant. With some trivial exceptions, any universal statement (i.e., any statement of the form "all *x*'s are *y*'s"; e.g., "all birds have feathers") is incapable of being definitively verified. (The trivial exceptions concern cases where the generalization concerns only some very circumscribed class of objects; e.g., "all pencils on my desk are yellow," or where some statement that, in terms of its content, is not a universal generalization but is turned into one through syntax-chopping; e.g., "all objects are such that, if they are identical with George W. Bush, they are mammals.") Since determinism is given by the statement "all events have complete causes," it is unverifiable, like "all metal expands when heated." On this basis, it is concluded that determinism couldn't possibly be known to be true, and shouldn't be affirmed, even tentatively.

But is it so clear that "all metal expands when heated" is unverifiable? So far as that is true, it's only in the trivial sense in which nothing other than analytic truths can be definitively verified. To be sure, "all metal expands when heated" couldn't possibly be verified on *strictly observation grounds.* After all, that statement concerns *all* metal, not just the metal you see. So no universal generalization, with the irrelevant exceptions noted a moment ago, can be verified on *purely* observational grounds; some element of inference is involved. *Anything* that can be observed to occur is, for that very reason, not general. And that is why, despite its syntactic similarity to "all metal expands when heated," "all pencils on my desk are yellow" is not *substantively* (as opposed to just syntactically) a universal generalization: one can *observe* that it is true. You can observe this or that metal bar expanding; but you can't observe the "general fact," if there even is such a thing, that heated metal always expands.

Also, if by "verification" you mean "observation-based verification," then "all metal expands when heated" and "every event has a cause" are incapable of verification, but only because "verification" has been so defined as to render that statement an empty truism.

In any case, if one allows at least some inference-rules to be permissible sources of verification, then "all metal expands" is verifiable, depending on

what those rules are. And if those rules are anything like those that laypersons and scientists permits themselves, then “all metal expands when heated” is verifiable. So it is only in some trivial sense that “all metal expands when heated” is unverifiable.

4.1 Why determinism, if true, is capable of being verified (continued)

Some say that “all metal expands when heated” is unverifiable in the sense that, no matter how many metal objects you’ve seen expand when heated, the next one “could” fail to expand. But what does “could” mean in this context? There is a sense in which I “could” win the lottery, or even sprout wings, even though the latter and probably also the former are prohibited by natural law. (In fact, the former is definitely prohibited by natural law since, given who I am, it isn’t psychologically possible for me to buy lottery tickets. I find the very concept of a lottery appalling.) To say that I “could” sprout wings is not to say that it’s permitted by natural law, let alone that it’s possible in some more robust sense. It is only to say that, given what I know, it isn’t ruled out. But I would contend that I do know that I won’t sprout wings and, therefore, that what I know does rule it out. And you too, outside of a philosophy class, would say that you know that you won’t sprout wings.

David Hume famously argued that inductive inference isn’t rational. His argument is not easy to refute. (That said, we saw in Chapter 8 how to refute it.) But nobody accepts Hume’s conclusion; and while it’s one thing to say that we don’t know how to justify a belief, it’s quite another to say that it has no justification. Surely my belief that I won’t sprout wings in the next five minutes is not in the same category as a person’s belief that tomorrow it will start raining gumdrops. And it’s a veritable truism that the difference is that the former is justified, whereas the latter is not.

4.2 An oft-heard argument as to why determinism, even if true, is incapable of being verified

Here is an argument that several students, independently of one another, have presented to me, and that is also found in the work of C.S. Peirce (1839–1914):

> Let cell *X* be a "perfect" clone of cell *Y*, and suppose that *X* and *Y* are kept under exactly similar conditions. (So, although they are kept in different Petri dishes, they are subjected to precisely the same forces.) *X* and *Y* will behave differently from each other. Therefore, the same initial conditions fail to lead to the same non-initial conditions. According to determinism, this isn't possible. (If two qualitatively identical circumstances are succeeded by qualitatively different circumstances, then neither circumstance completely determined what it would be succeeded by.) From this it follows that determinism is false.

Given only that two cells are biologically exactly alike, it doesn't follow that they're exactly alike in all other respects. It doesn't follow, in particular, that they are exactly alike at the atomic and sub-atomic levels. And the chances of their being thus alike are basically zero. There is no real chance that the micro-particles composing one organism will be moving in precisely the same way as the micro-particles composing any other organism. These sub-biological facts can and will have dramatic effects on what happens at the biological level. All this argument establishes, then, is the irrelevant point that *different* initial conditions may lead to *different* non-initial conditions.

4.3 Another oft-heard argument as to why determinism, even if true, is incapable of being verified

Many students of mine have put forth the following argument:

> Determinism says: "same initial conditions, same non-initial conditions." But the conditions obtaining at any given time are different in at least some respects from those obtaining at any other time. For, between any two points in time, there has surely been at least one change; some particle has moved or rotated or vibrated.[303] So there is no way to verify determinism.

Causal connections are discovered by seeing how differences in initial conditions correlate with differences in non-initial conditions. You allow initial conditions to change in *one* respect, but do your best to ensure that they're unchanged in every other respect. This will guarantee that these other factors will at least *tend* to remain constant, even though, inevitably, they'll never be perfectly replicated. If changes in that one factor reliably correlate with changes in the outcome, there's probably a causal connection; and there's probably no such connection if there's no such correlation.

You may not be able to establish a *perfect* correlation between your pushing the button and the elevator's coming. But that doesn't mean that you can't have good reasons for positing the existence of a strictly deterministic connection between the first and the second. For there to be such a connection is *not* for the elevator to come without fail every time the button is pushed. It is for the likelihood that the elevator will come, given that the button was pushed, to approach 100% to the extent that other factors are held constant; that is, it is for there to be no *limit* to how close to 100% that likelihood can be made to be so by rooting out other factors that aren't constant.

Your experience can easily give you good evidence that *this* condition is satisfied. If you find that it's always possible to make it *more* likely than before that the elevator will come, given that the button was pushed—if, in other words, you find that, given enough tinkering about with the relevant circuitry and cables and so on, it's always possible to tighten this correlation—you *ipso facto* have evidence that the condition in question is satisfied and,

therefore, that a strictly deterministic connection is at work.

This doesn't show they can be definitively established. But it *does* show that, insofar as they're incapable of being definitively established, it's only for the reason that *nothing* known through sense-perception can be definitively established. Your sense-perceptions make it probable that you have a television, but they don't make it certain. They make it probable that you're not a disembodied spirit, but they don't make it certain. They don't make anything about the external world certain. A fortiori they don't make the existence of any causal connections certain. But such connections are no less capable of being definitively established than anything else about the external world. Whatever shortcomings induction may have as a form of inference, they're no more of a problem for attempts to establish causal connections than they are for attempts to establish that you have hands.

We can't have perfect evidence of a perfect correlation.[304] But that doesn't mean we can only have evidence for imperfect correlations. It's possible to have imperfect evidence for perfect correlations. I've examined a million mammals, and they all had backbones. Here we have evidence of a perfect correlation. It's imperfect evidence, but that's the only kind that's available. More precisely, it's imperfect evidence *if* by "perfect evidence" we mean definitive *observational* evidence—as opposed to observational evidence *plus* reasonable inference-rules. But if this is what we mean by "perfect" evidence, then it's a sterile triviality that we can't have perfect evidence for the existence of transperceptual correlations. (By "transperceptual," I mean "happening on the other side of the veil of perception—out there in the world, as opposed to in our minds.")

We have similarly "imperfect" evidence for deterministic laws. And such laws are no *more* incapable of being definitively established, and no less capable of being established with reasonable certainty, than biological concomitances of the kind just described.

5.0 Some general, but often overlooked, principles concerning determinism

A deterministic *system* may consist of parts that behave indeterministically. This is a consequence of a more general principle, to wit: the laws governing the parts don't necessarily govern the whole. (In Chapter 19, we'll see that attempts to reduce psychology to physiology involve a failure to grasp this principle.)

A story may help make this clear. General Smith is in charge of Battalion A. The members of *A* are extremely disciplined, and none of them would dream of disobeying the command of a superior officer. Smith's commands determine how *A* as a whole will behave. But those commands don't micromanage every detail about every soldier's behavior. They don't determine the order in which Private Jones puts on his boots or exactly how the muscles in Private Brown's mouth move when he's eating. The forces that determinize (i.e., render deterministic) the behavior of *A as a whole* don't determinize the behavior of *A*'s constituents.

It may be that *other* forces determine the behavior of *A*'s constituents. But, even if this is true, we're seeing some facts of profound importance. A system cannot *just* be deterministic; it must be deterministic *in this or that respect.* It makes no sense to say, without qualification, that a system is *just* deterministic---unless by that it is meant that it's deterministic in *every* respect.

Also, we're seeing that the forces in terms of which the behavior of a *system* is to be understood don't necessarily coincide with those in terms of which the behavior of its parts are to be understood. So supposing for argument's sake that our thoughts are brain-states (or changes therein) and, consequently, that they consist of microparticles, it doesn't follow that psychology is a branch of microphysics; it doesn't follow, in other words, that the principles in terms of which a person's thoughts and feelings (and so on) are to be understood coincide with those in terms of which the microparticles composing them are to be understood.

In objection to this last point, it may well be said that:

> (MP)[305] You're overlooking an important fact. The reason that Battalion A's behavior is determined by what General Smith says is that each individual soldier in *A* knows that his

> career would be over if he were to disobey the command of a superior officer. Were it not for Private Brown's knowledge that he'd be court-marshaled if he decided to flout a C.O.'s orders, Brown wouldn't currently be hiking through a mosquito-infested jungle wearing a 70 pound pack. So, at the end of the day, it's the micro-mechanisms that do all the work.

True. It *is* the microstructures that do all the work. It's always the microstructures that implement the "higher-level" forces. For this reason, people advocate MP; in other words, they conclude that these higher-level forces don't really exist and also that everything that happens is to be explained and understood in terms of micro-(re)distributions of mass-energy. But they're wrong. An extension of our story will make this clear

Battalion A is but one of the many battalions that General Smith commands. Smith also commands Battalion B. *A* and *B* have the same structure: for each Captain in *A*, there is a Captain in *B*, and *vice versa*; and given any Captain in *A*, if that Captain is a fighter-pilot, so is his counterpart in *B*. *A* and *B* are perfect isomorphs of each other.

But at the *sub*-structural level, *A* and *B* are very different. The members of *A* tend to be intellectual. Some of them are athletic, but their preferred sport is badminton. The members of *B* are anti-intellectual brutes; their favorite past-times are getting drunk and throwing sandbags at each other. But because of the way in which the members of each battalion are *interrelated*, these differences have no effect on how either Battalion acts *as a whole*. The principles in terms of which those Battalions behave, when taken as wholes, is to be understood in terms of principles that won't be apparent, at least not for what they are, at the level of this or that particular soldier's personality or conduct.

> But surely if you knew exactly what each soldier in *A* was doing, and you knew exactly what his counterpart in *B* was doing, you could, in theory, deduce how *A* and *B* were behaving. And, by obvious extensions of this reasoning, you could easily explain the fact that *A* and *B* are, as wholes, qualitatively identical in terms of lower-level forces; i.e., in terms of the forces that govern the parts of *A* and *B* (and, ultimately, the parts of the parts [of the parts of the parts…] of

the parts of *A* and *B*).

I agree. But these correct points leave out some truths that, given the context, shouldn't have been left out. *Given* a knowledge of how (say) *A* behaves as a whole, one can *post hoc* see that *A*'s behavior was strictly determined by the behavior of its parts. (I'm supposing for argument's sake that acquiring the relevant information about *A*'s parts won't be a problem.) That micro-information would have been explanatory of *A*'s behavior as a whole *only* to the extent that given similarly structured entities, whose parts may be very unlike *A*'s parts, it *also* explains *their* behavior under circumstances similar to those in which *A* was operating.

5.1. An illustration of this principle

A story may clarify my meaning. There is some time *t* such that you know exactly what each member of *A* is thinking and doing at *t* and such that you know exactly what each member of *B* is thinking and doing at *t*. (How you know this isn't part of the story, since it's irrelevant to what we're talking about.) On this basis, you figure out what each of *A* and *B*, taken as a whole, is doing. So, at *t*, *A* and *B* are behaving in exactly the same way---despite the fact that, at *t*, there isn't a single soldier in either battalion who is acting just like his counterpart in the other battalion. Any battalion whose members are *interrelated* in the way in which, at *t*, *A*'s members were interrelated will, under qualitatively identical circumstances, behave exactly as *A* behaved (and, therefore, exactly as *B* did). So supposing that *A* has organizational properties $P_1 \ldots P_n$, anything else that has those properties will react to any given situation in exactly the way that *A* would, even if that other thing's parts are very different from *A*'s parts.

So, when asked "why did *A* do such and such at *t*?" the right answer is *not*: "because Private Green was doing such and such, and Major Max was doing thus and such…" That answer would leave out the information about *A*'s *structure* that guaranteed that anything having that structure would have acted in the same way had it been in the same circumstances. That answer would have provided irrelevant information about the specific way in which that structure was implemented on this particular occasion.

Micro-knowledge is knowledge of how the forces governing macro-structures are *implemented.* But it isn't knowledge of those forces themselves. To be sure, everything that happens at the macro-level is fixed by what happens at the micro-level: it is impossible, even in principle, for there to be two worlds that are micro-identical but macro-distinct. But it doesn't follow that the macro-world is not to be understood on its own terms; it doesn't follow, in other words, that the macro-world obeys laws that, although implemented by events governed by micro-laws, are altogether distinct from those laws. Thus, it doesn't follow that the truths provided by disciplines other than microphysics are merely low-grade microphysical truths. In Chapter 19, we'll find that the general principles about the structure of the world of which such a science apprises us are categorically incapable of being interpreted as (low-resolution) microphysical truths. At the same time, we *will* find that *specific* truths belonging to higher-level physical

sciences (e.g., botany, astronomy) are capable of being interpreted as low-resolution microphysical truths. But, finally, we'll *also* find that psychological truths, whether general *or* specific, are categorically incapable of being interpreted as truths of microphysics or indeed of any science. At the same time, we'll see (in Chapter 22) that psychological truths are almost certainly identical with the (re)distributions of mass-energy that fall within the scope of microphysics. It will be made clear in Chapter 21 how all of these facts can be reconciled with one another.

5.2 Ramsey's conception of physical law: an illustration of the vicious circularity inherent in hyper-empiricism

A number of authors (e.g., Wittgenstein, Ayer, Ramsey, Davidson, Carnap) have said, in characterizing the world as "deterministic," one isn't directly saying anything about that world. Rather, what one is saying is along the lines of:

> (AR)[306] Let $T_1 \ldots T_n$ be a list of truths that jointly describe everything that happens throughout the course of time. Let *T* be this giant list of truths. There will be ways of organizing *T* that enable you to compress it. There will be *general* propositions---propositions having such forms as:
>
> (*) anything that has phi also has psi,
>
> and:
>
> (**) if there's an occurrence in a given space-region of an event having phi, there will be an occurrence in a neighboring region of an occurrence having psi,
>
> and the like.
>
> Your objective, let us suppose, is to compress *T* as much as possible. In other words, you're trying identify some truth *T** such that:
>
> (i) *T** is much shorter than *T*, the reason being that there many things that *T explicitly* says that *T** does not say explicitly,
>
> but such that:
>
> (ii) *T** contains all the information contained in *T*, the reason being that, for any truth that *T* but not *T** states explicitly, *T** states it *implicitly*----in other words, that truth can be *deduced* from the truths in *T**.

Some of the truths on *T** will concern specific matters of fact---they will describe how, for at least one time *t*, mass-energy is distributed at *t*. (Given the points earlier made about instants, what this really means is that, for at least one time-*period t*, *T* will say how mass-energy is (re)distributed during *t*.) But it will provide as few truths as possible. The remaining members of *T** be will *general* propositions that, *given* the remaining truths on *T**, enable you to deduce all the truths that are not on *T** but *are* on *T*.

Let's suppose that *T** is the most compact list of its kind. In other words, it's the shortest list of truths on the basis of which everything that is on *T* can be deduced. The general propositions on *T** are what we call "the laws of nature."

Thus, the laws of nature are just abbreviatory devices. They spare us the trouble of having to know various individual facts; we need only know a few individual facts, along with a few general principles; and we can figure out anything else that we need to know as the occasion arises.

There may be distinct, equally compact ways of compressing *T*; and given two such shortenings of *T*, they may not contain the same general propositions. That doesn't mean that one of them is wrong. It shows that the laws of nature aren't "out there," in the world, and are rather "in here," in our methods of describing the world.[307]

Here's a Mickey Mouse example of the principle at work here. Let *L* be a list that says of each individual person how many cars he or she has. Now imagine adding another list to *L* that gives that same information, but does so indirectly: instead of saying "Bob has five cars," this second list says "Bob has 3 + 2" cars. Now imagine a third list that gives that same information but in some *third* way. This third list says "Bob has the square root of 25 cars." And let's suppose that we take 20 more lists such that:

(a) no two of them are identical with each other or with any of the lists just mentioned,

(b) each says how many cars any given person has, but

(c) each does so by providing an expression that can be proven, on the basis of arithmetic, to refer to the relevant integer.

Really, *L* provides no less information than this super-list. If you know that Bob has 2 cars, you don't need to know that, for some number *n*, Bob has *n* cars and $n^5 = 32$. But, given this super-list, in order to see that it gives you the same information as *L*, you need knowledge of arithmetic. So the laws of arithmetic serve an abbreviatory function. Laws of nature are to be understood in much the same way. They are to be seen as abbreviatory devices, not as forces or structures that regulate the course of events.

5.2.1. The spuriousness of AR and the uselessness of the Ramsey-sentence

References to the laws of nature, if only implicit ones, are an ineliminable part of any description of any event. This is because, contrary to what AR assumes, space and time aren't vessels that happen not to be empty but could be.

The concept *occupancy*—as in occupying a certain space---is a causal concept. Smith wants to sit in chair *x*. By virtue of that Jones is already sitting in *x*, what must happen in order for Smith to sit in *x* will be very different from what *ceteris paribus* would otherwise have to happen. Even if Smith just walks over to *x* and sits down in it, Jones' body will serve as a buffer between him and *x* itself. It makes no sense to say that, while Jones is sitting in *x*, his presence there has consequences as regards what must ensue if, in the next two seconds, Smith is to be sitting in *x*. It makes no sense to say that, even though Jones is sitting in *x*, light is passing through the region he is occupying in exactly the way it would pass through that region were it unoccupied. And so on.

Spatiotemporal relations are causal relations. (See Chapter 12 for a lengthy discussion of this.) It would make no sense to say that John's flipping the switch caused the light to turn on, but that the very instance of that light's turning preceded that very instance of John's flipping that switch. If the latter caused the former, it *ipso facto* preceded it. According to Relativity Theory, if a light-beam, sent by event *x* at the time of its occurrence, reaches *y*, just as it's beginning to occur, then relative to any possible frame of reference, *x* precedes *y* in time. If a light-beam sent by *x* doesn't reach *y*, *and* a light-beam sent by *y* doesn't reach *x*, then *x* and *y* are simultaneous with each other relative to some frames of references and, relative to others, *x* precedes *y*, and relative to others, *y* precedes *x*. But, according to Relativity Theory, the causal order doesn't merely *reflect* the spatiotemporal order; it *is* the spatiotemporal order.

A cause is a law at work. Setting aside exceedingly low-resolution descriptions of events, there is no description of an event that is nomically antiseptic. AR assumes that descriptions of the world can be neatly segregated into two groups, ones consisting of propositions that express laws, the other consisting of propositions none of which makes any assumptions as

to what laws obtain, if any. But that's not true. Once you set the facts, you *ipso facto* set the laws. The reason lies in the following two facts, taken jointly:

> (1) You can't say whether (relative to a given framework) A preceded, followed, or was simultaneous with *B* until you establish *either* that the one event is part of a causal series beginning with the other, in which case the latter was the first to occur, *or* that there is no such series, in which case they're simultaneous.
>
> (2) Wherever there's causation, there's law. If e_1 caused e_2, then there is some law *L*, and some properties phi and psi such that e_1 has phi, such that e_2 has psi, and such that *L* is given by a proposition to the effect that anything having phi (occurring in a given region) precedes something having psi (occurring in some region related, in some specified way, to the first region).

To sum up: AR naively assumes that there are *anomic* facts. ("Nomic" means "having to do with natural law. ""Anomic" means "free from the influence of natural law and therefore capable of being understood independently of anything having to do with natural law.") In fact, in describing something as "pinewood" or as "copper," we're assuming that certain laws hold with respect to it. If it doesn't conduct electricity, is it copper? If it looks just like solid oak, but you can walk right through it, is it really solid oak?

The terms that we use to pick out the objects that participate in the events and facts described by *T* are nomic terms. In describing something as copper, or water, or human, you are, in so doing, ascribing to it certain causal properties.[308] (You may not know what those causal properties are. There are certain properties $P_1 \ldots P_n$ such that, if I say that *x* is a puddle of mercury, the chemist to whom I am speaking will automatically know that, if my statement is correct, *x* has $P_1 \ldots P_n$, so that, in effect, I was in *some* sense ascribing those properties to *x*. A derivative sense? Yes. An unimportant one? No.) Since AR assumes not just that some events can be understood in thoroughly anomic terms, but that all of them can be so understood, AR isn't

even remotely close to the truth.

6.0 The concept of measurement in relation to the concept of an instant

Like the expression, “the average person," the word “instant” refers to nothing, its function being to make it possible to abbreviate what would otherwise be long and cumbersome statements. There is no human being *x* such that the sentence:

(AP)[309] “the average man owns 2.2 cars,”

which says of *x* that he owns 2.2 cars. AP is an abbreviation of:

(AP*) If *x* is the total number of cars that some person or other owns, and *y* is the total number of people, then $x/y=2.2$.

Neither “the average person” nor any synonym thereof occurs in AP*. Indeed, there is no isolable part of AP* that corresponds to the occurrence in AP of “the average man.” That expression refers to nothing. But “the average man” is nonetheless meaningful. So, given that its meaning isn’t given by saying what it denotes, how *is* its meaning to be given?

6.1 The word “instant” to be defined contextually, not denotatively

Here a brief detour through the philosophy of language is in order. “The average man” is in the same grammatical category as expressions that *do* refer to specifics; e.g., “the current U.S. President” and “the Earth’s only moon.” For this reason, a sentence of the form “…the average man…” *ipso facto* appears to concern some individual; it appears, to be more precise, to say of some individual *O* that….*O*…(i.e., that *O* is a thing *x* such that…*x*…) Since that is *not* what is meant by any given sentence *E* of the form “…the average man…” and since any sentence *E* that states *E**’s meaning in a non-elliptical manner has a grammatical structure that is to no degree comparable to *E*’s, it follows that *E*’s grammatical structure is at variance with its meaning. Thus, in virtue of having the form “…the average man…” a sentence’s grammatical form is out of line with its logical form. The meaning of “the average man” is given *not* by saying what it denotes, but by saying how, when given an arbitrary sentence *E* of the form “…the average person…” to identify a sentence *E** such that:

> (i) *E* and *E** say the same thing, and
>
> (ii) *E**’s grammatical form doesn’t misrepresent its logical form in the way in which *E*’s grammatical form, *by virtue of the fact that it contains an occurrence of “the average man”* misrepresents *its* logical form.[310]

Everything just said about “the average man” is true of the word “instant.” There are no instants. If *E* is an arbitrary sentence of the form “…instant…” *E* is but an abbreviation for some sentence *E** that does not contain that word or any isolable part corresponding thereto. The meaning of the word “instant” is not given by a rule of the form: the word “instant” refers to entities having such and such properties. Rather, it is given by a rule to the effect, if *E* is a sentence of the form “such and such is the case at a given instant *t*,” *E* is an abbreviation for some sentence *E** of the form[311]:

> (R_1) “For any *t*-inclusive non-null interval S_1, if the world’s condition during S_1 deviates from such and such by amount *M*,

> there is some smaller *t*-inclusive interval S_2 such that the world's condition during S_2 deviates from such and such by less than *M*."

"R_1 is viciously circular and therefore fails," it will be said, "as it defines 'instant' in terms of itself, given that a '*t*-inclusive' interval is one that includes *instant t*."

In Sections 7.5-7.5.3, it will be made clear why R_1 is not viciously circular. For now, let us operate on the (soon to be discharged) assumption that R_1 is non-circular and otherwise acceptable.

6.2 The relevance of these semantic points to the concept instantaneity

One cannot directly determine how quickly the car is moving *at* instant *t*. What is it to determine how fast the car is moving during some interval? First of all, measurement is comparison. Given some object *O*, to measure *O*'s length is to determine how long it is *relative* to some object that is taken as a standard. (In a moment we'll discuss the criteria that an object must satisfy to qualify as such a standard.) It is determined that, for some object *O** such that *O**'s length is by definition one unit long, and for some number *n*, *O*'s length is n multiplied by *O**'s length. To determine *O*'s mass is to determine how *O*'s mass compares with that of some *O** such that *O** by definition has a mass of one; it is to determine, for some number *n*, that *O*'s mass is *n* multiplied by *O**'s mass.

All magnitudes are relative; some objects are longer than others; some objects are more massive than others.[312] In order for a given object *O** to be such that it can appropriately be taken as a standard, it is necessary and sufficient that that there be no object *O#* such that, if *O#* is taken as a standard, there are fewer causal anomalies than there are if *O** is taken as a standard.

Let *O** be some inflatable raft that we stipulate to be one unit long, no matter what changes it undergoes. A consequence of this stipulation is that inflating *O** would have the consequence that, apart from *O** itself, *every* object in the universe shrinks, even though:

> (i) many of those objects are on the receiving end of any causal process initiated by *O**'s being inflated,

and even though:

> (ii) given virtually any object that is on the receiving end of processes initiated by *O**'s being inflated, it's shrinking has nothing to do with that fact about it.

Object-shrinkages of the first kind would be uncaused and, therefore, anomalous. Although they'd be consequences of *O**'s being inflated, they would be *logical*, not *causal*, consequences of *O**'s being inflated. So they'd

be uncaused events. They'd be anomalies.

Object-shrinkages of the second kind would also be uncaused. It is irrelevant that those objects were on the receiving end of causal processes initiated by *O*'s being inflated. The laws of physics being what they are, *O**'s being inflated couldn't possibly have caused those objects to shrink.

Also, *O**'s being inflated would have the consequence that distances between objects became smaller. This consequence, it goes without saying, would be quite as inexplicable as the two just mentioned.

Of course, in order to validate our decision to take *O** as having an unchanging length, we might posit physical laws that would explain why *O**'s being inflated led to such consequences. But those physical laws would themselves be inconsistent with well-established laws (e.g., that radiation travels at a finite speed), and we couldn't eliminate those well-established laws from our system of physics without ceasing to understand what we previously understood and without also ceasing to be *capable* of understanding both what we already understood and what we did not yet understand but were capable of understanding.

To sum up, a consequence of taking *O** as having an unchanging length would be that, to a large extent, the world would cease to be intelligible to us. Given *any* object that we intuitively regard as being highly alterable in some respect, an analogue of this argument shows, were that object stipulated to be unchanging in that respect, the results would be comparably devastating to our ability to understand the world.

6.3 The relevance of these semantic points to the concept instantaneity (continued)

Now suppose that one takes as one's yard-stick, as it were, some titanium object that is kept in an environment in which, speaking intuitively, it doesn't undergo any (relevant) changes---that it is kept in a temperature controlled room, that it isn't subject to forces that would disfigure it, etc. Taking *that* object as having an unchanging length *won't* lead to anomalies of the just-discussed kind and *won't* strip the world of whatever causal and explanatory integrity it has.

6.4 How measurements of speed are, in the relevant respects, comparable to measurements of mass

Speed, like mass and volume, is to be understood in relative terms. Consider the statement:

> (1) During time-interval *T*, *x*'s average speed is such and such.

The meaning of (1) is to be understood in terms of the concept of a *periodic process*. A periodic process is a repetitive sequence of events. More precisely, it is a series of events each of which is of the same kind as its neighbor(s) in that sequence (all of those members have two neighbors except for the terminal members). The series of rotations of the Earth around the Sun, during an interval long enough for more than one such rotation to occur, is a periodic process, and so, during such an interval, is the series of rotations of the Earth on its axis. A person's heartbeat is a periodic process, and so is the series of round-trips made by a beam of light bouncing between two mirrors whose distance from each other remains constant.

In some cases, *direct* comparisons of speed are possible. If *x* and *y* are engaged in a foot race, then if *x* makes it to the finishing line before *y*, *x*'s average speed during the interval when *x* was running exceeded *y*'s average speed during that same interval. (We are assuming that, for the duration of the race, *x*'s trajectory was the same as *y*'s and, therefore, that neither covered a significantly greater distance than the other.)

6.4.1 How measurements of speed are, in the respects that are relevant here, comparable to measurements of mass (continued)

But direct comparisons of speed are not always possible. Suppose that runners *x* and *y* both travel from points *A* to *B*, whose distance from each other is held fixed, but that they do so at different times. *x*'s trajectory, we will assume, is the same as *y*'s.

Given those circumstances, how is it to be established who ran faster? One must first identify some periodic process; e.g., a light-beam bouncing between two mirrors held at a fixed distance from each other or ticks on a properly functioning clock. If fewer installments of that process occur during *x*'s journey from *A* to *B*, then *x*'s average speed during that journey exceeded *y*'s.

Bearing this in mind, suppose that the periodic process in question is one that, speaking intuitively, we'd describe as highly variable. Suppose that Smith's heart rate is this process and that Smith suffers from a heart-problem that involves his suffering from both extreme tachycardia and bradycardia. (Speaking intuitively----in other words, speaking as though it made sense to speak of absolute speeds---"tachycardia" and "bradycardia" may be defined, respectively, as "hyper-accelerated heart rate," as when one is having a heart attack, and "hyper-decelerated heart-rate," as when one is in a coma.) In that case, even if *x* is an Olympic sprinter who is performing at his best during his journey from A to B, and *y* is an arthritic who couldn't even make it onto his school track team, Smith's heart may beat more times during *x*'s journey than during *y*'s.

Let us generalize this point. Suppose that we take Smith's heart-rate as being unchanging—that, by our stipulation, it doesn't change and that, serving as it does as a "standard" process, we determine the speeds at which objects are traveling in terms of it. In that case, we'll end up with counterintuitive results, such as the one just described, and we'll also end up with rampant anomalies, comparable to those described earlier in connection with mass and volume. Every time that Smith has a heart-attack, everything in the universe will slow down; and every time Smith falls asleep, everything will speed up. For reasons analogous to those described earlier, this will have two consequences:

(i) There will be otherwise non-existent causal anomalies,

and:

> (ii) There will be (otherwise non-existent) nomic anomalies (i.e., there will be laws of physics that are inconsistent with other, better corroborated laws of physics).

To sum up, by stipulating to be non-erratic a periodic process that we intuitively regard as erratic, we strip the universe of much of its causal and explanatory integrity.

6.4.2 How measurements of speed are, in the relevant respects, comparable to measurements of mass: closing the argument

But now suppose that, given some period process that we *do* intuitively regard as non-erratic (e.g. the series of round-trips made by a light-beam between two mirrors whose distance from each other remains fixed), we stipulate that it is indeed non-erratic. This practice will have none of the just-mentioned consequences; it will *not* have the result that Olympic sprinter x is a slower runner than arthritic, non-sprinter y; it will *not* have the result that the rate at which a given object O is traveling will change depending on the vicissitudes undergone by some object O* to which O is not (relevantly) causally connected. By the same token, if we stipulate that a process of the just-discussed kind *is* erratic, the consequences will, by now familiar reasoning, resemble the consequences of stipulatively regarding as non-erratic some process that, intuitively speaking, *is* erratic.

Thus, given two periodic processes *P* and *P**, *P* is less erratic than *P** iff *ceteris paribus* there are fewer causal and nomic anomalies if *P* is taken as a standard than if *P** is taken as a standard. Thus, a given periodic process *P* is *perfectly* non-erratic if there is no periodic process *P** such that *ceteris paribus* fewer anomalies would result if *P** were taken as a standard than if *P* were taken as standard.

6.5 Why R_I is not viciously circular

R_I may initially *appear* circular. But it appears that way only so long as one holds onto the demonstrably false idea that position in time is non-relational. Consider the sentence:

(S) "Event *x* occurs at 3:00 P.M. on April 21, 2010."

S says that *x* occurs simultaneously with some event *y* such that *y* is an installment in a certain kind of periodic process that is assumed to be unchanging, the reason being that not making this assumption leads to a proliferation of anomalies that would otherwise be easily explained. To say that JMK sneezed at a given time is to say that his sneezing was simultaneous with, let us say, the coincidence of some needle on some clock with some mark on that clock's face or, more realistically, with the collision of some light-ray with some surface.

"But," some will say, "to say that two events are 'simultaneous' is to say that they occur at the same instant; so your analysis is still viciously circular."

Not true. To say that two events are simultaneous is to say that the laws of physics prohibit either event's having an effect on, or a hand in, the genesis of the other. (For an explanation of this statement see Chapter 19, Section 8.0.) Thus, the statement "*x* and *y* occurred simultaneously"[313] means that *x* cannot have any effect on *y* or *vice versa*.[314] Given an arbitrary event *e*, let *E* be the smallest class containing all and only events *e** such that *e* cannot affect *e** or be affected by it. In that case, the statement "*x* is a member of *E*" is equivalent with "*x* occurred at the very same instant as *e*" or, alternatively, "there is some instant *t* such that *x* and *e* both happened at *t*." Thus, an instant is the least inclusive class of events all of whose members are incapable of affecting or being affected by some given event; and an event's happening at that instant is identical with its being a member of the aforementioned class. Given any class *C* such that, for some event *e*, *C* is the least inclusive class containing *e* and every event that *e* can neither affect nor be affected by, let us say that *C* is an *I*-class.

Instants are *I*-classes. There is no way to pick out an *I*-class except by picking out one of its members. The reasons for this are of a strictly logical

nature; they have nothing to do with anyone’s limitations.[315]

7.5.1 Why R_I is not viciously circular (continued): A penultimate preliminary point

An interval of time is a class *C* of *I*-classes such that:

> (i) The (distal) cause of any event that is a member of the last member of *C* is a member of the first member of *C*.
>
> (ii) if *x* is any given member of *C*'s terminal *I*-class and *y* is any given one of the members of *C*'s initial class that is causally responsible for *x*, then, for each event *z* such that *z* was one of the events that was a link in the chain of events whereby *y* affected *x*, there is some *I*-class that belongs to *C* to which *z* belongs.

Given any series of *I*-classes satisfying these requirements, let us refer to them as *S*-classes.[316]

7.5.2 Why R_I is not viciously circular (continued): final preliminary point

There is one last point to make before closing the argument. We assign numbers to instants. The higher the number, the later the instant. So "*t*" is a variable, the constants corresponding to which are numbers. There must be some *one* relation *R* such that if *e* is a member of some *I*-class *X*, that *I*-class must be assigned a number that bears *R* to that assigned to any *I*-class to which one of *e*'s effects belongs. (We use the relation "lower than," since it has the right properties and since, unlike some of the other relations that have the right properties, it's psychologically easy to manage. So if *e* (somebody's birth) happens before *e** (that person's beginning to sneeze on his second birthday), then *e* is assigned a lower number (say, 06:45, June 8,1970) than the one assigned to *e** (09:45, June 8,1972).

Each event is (in principle) assigned a number. But the number it is assigned is *not* based on the "position" in time that it occupies. There is no such position. It is, since it must be, based on its causal relations to other events.[317] *e* is given a higher number than *e*# if *e* is incapable of affecting *e*#; *e* is given the same number as *e*# if neither can affect the other. Given some event *e*, our conventions assign it a number on the basis of a procedure consistent with the guideline just described; and given one's knowledge as to the identities of the numbers assigned by the same procedure to other events, one knows (within limits) what is (simultaneous with something that is) the effect of what.

7.5.3 Why R_{I} is not viciously circular: closing the argument

Bearing all of this in mind, let's revisit the question: what does it mean to say that a given time-interval is "*t*-inclusive"? *t* is an *I*-class. *t* is assigned a number. (Or, what comes to the same thing, each member of *t* is assigned a number, and that number is then assigned to *t* itself.) For argument's sake, let 34 be that number. *t* is a member of an *S*-class, which we'll call "*X*." This means only that any given member of *t* is either the cause (not necessarily the immediate cause) or the effect (same qualification) of at least one member of at least one of *X*'s terminal *I*-classes.

So to say that a given interval is *t*-inclusive is just a very encrypted way of saying that a certain event (it could be a bolt of lightning, a hiccup, or the ringing of an alarm clock) is a member of a class no two of whose members can affect each other but any given one of whose members is the cause or the effect one of the events marking either that interval's inception or its cessation. It follows that occurrence of "*t*-inclusive" in R_{I} is eliminable, being but an abbreviated way of making a causal claim of the kind just described.

Chapter 14 Causality

1.0 The basics

Events cause other events to occur. Smith pulls the trigger, causing the gun to fire. Jones throws the vase against the wall, causing the vase to break. There are causes and there are effects.

Astonishingly, some philosophers have denied the existence of causation. Why? Because many philosophers are ***empiricists,*** and given an empiricist outlook, it is hard to accommodate the presumption that events make events happen. According to empiricism, all knowledge is strictly observation based. Thus, if you know it, one of two conditions is met: (a) you observed it (in other words, you saw it, heard it, felt it, etc.); or (b) you inferred it from what you observed, *provided* that you used no inference rules other than those of whose legitimacy direct observation apprised you. (See Chapter 12.) According to many empiricists, we cannot see or otherwise observe causal ties. We see events, but not forces. We see rocks strike windows, and we see windows shatter upon being thus struck. But we don't see anything that, given such a collision, *compels* the window in question to shatter. Since we don't observe causal connections, we don't have good reason to believe them to exist.

But this argument involves a major non-sequitur. We don't observe electrons or quarks, but we have good reason to believe them to exist. How are causal ties different? Given that it's methodologically permissible to posit unobservable viruses and particles, why isn't it equally permissible to posit unobservable causal ties?

"Because," we are told, "the concept of a causal tie, unlike that of a proton or a virus, is an incoherent one. For that reason, it wouldn't do any good to posit causal ties. Causal ties, if they exist, are occupants of space-time, like rocks and vases and chairs; and they are therefore just as incapable as chairs and rocks of being the "glue" that, supposedly, keeps the universe from powdering into so many renegade events." But notice that the second argument, unlike the first, isn't an empirical argument at all. It's a strictly logical one. It is to the effect that forces are in the same category as square circles and therefore couldn't *possibly* exist and, therefore, couldn't possibly be observed. Thus, the first empirical argument against the existence of causal ties is parasitic on a non-empirical, purely ***a priori*** argument, and is

therefore only as good as the latter. (An *a priori* argument is one that has its basis, not in observational data, but entirely in considerations of coherence and logic.)

Moreover, the latter argument isn't a good one. The concept of *persistence* is itself a causal notion, and anything that persists is, for that very reason, a causal sequence. Thus, we observe causal sequences whenever we observe anything. And we *knowingly* observe them. In other words, we don't merely observe what are *in fact* causal sequences; we observe such sequences *to be* causal sequences. This line of thought will be substantiated in Section 5.3.

Much of this chapter will be spent evaluating the contention that we don't see or otherwise observe causal ties. Although it has been enormously influential, it's of doubtful accuracy, as we've just seen. And the more it's scrutinized, the more incoherent it proves to be. For example, the very notion of spatial *occupancy* is a causal notion, since a given thing's occupying a given region *R ipso facto* alters the trajectories and velocities of other would-be and soon-to-be occupants of that region.

These points generalize. A "steel" object that lacked the requisite degree of malleability or thermal conductivity would *ipso facto* not be steel. Malleability and conductivity are causal properties. *x* is more malleable than *y* if, other things being equal, less force is needed to cause *x*'s shape to change than is needed to cause *y*'s shape to change. *x* has a higher degree of thermal conductivity than *y* if, other things being equal, less energy is needed to raise *x*'s temperature by a given amount than is needed to raise *y*'s temperature by that same amount. Thus, to describe a given object as being of made of steel is to say that, other things being equal, it will react to situations in ways in which it wouldn't otherwise react.

In describing something as "self-identical" or as "being such that 1 + 1 = 2," one is not making a causally loaded claim. But those characterizations say nothing about anything. In being told that *x* is "self-identical," one isn't being told how *x* is different from, or similar to, water or steel or the planet Neptune or anything else. The same is true of all causally sterile characterizations. For that reason, anything that says anything about anything in the spatiotemporal world has causal content.

2.0 Causation a relation between events, not objects

But before we enter these stormy and controversial waters, let us make some non-controversial points about the logic of the concept of causality.

First of all, it is only *events* that can be causes. *Objects* are not causes. The rock is not a cause. The rock didn't cause the window to break. What did so was some event involving the rock (e.g., the rock's striking the window at a certain time).

Sometimes the way we speak suggests that, in our view, objects are causes. For example, people say "Carter was the cause of the recession in the late '70s," "Hitler was the cause of World War II." But these statements are elliptical for "such and such events involving Carter (Hitler) were responsible for the recession in the late '70s (World War II)."

In connection with this, we must distinguish between something's being a cause, on the one hand, and its having causal properties, on the other. Non-causes may have causal properties. The rock isn't a cause, as we discussed, but it has causal properties. The rock's having a certain mass is among its causal properties, meaning that its having that mass is, or can be, a (partial) cause. (Not every property of the rock is a cause. The rock is such that it either is or is not identical with the number 2. But the rock's having that property doesn't cause anything to occur.)

2.1. Conditions vs. events

We must also distinguish conditions from events. A condition is a relatively persistent state (e.g., the room's having a certain temperature). An event is a *change* in conditions. But conditions *themselves* seem to be causes. For example, the water's having a certain temperature is what caused the spaghetti in it to soften. At the same time, conditions appear not to be events. An event is, by definition, a change, whereas a condition is, it would seem, the absence of change. So if conditions are non-events and are also causes, then, contrary to what we earlier said, it isn't events alone that are causes.

This argument overlooks the fact that conditions *consist* of events. The water's having a given temperature *consists* in its containing particles that have certain kinetic properties. There is no heat where there are no events. A person's being healthy (i.e., being in a condition of health) consists in that person's body doing what it should. There is no life, and therefore no health, where there are no metabolic events. And although events do necessarily constitute changes, those changes may be, and often are, internal to conditions of changelessness. For the water's temperature to remain at 101° for 5 minutes is for the changes undergone by its constituent molecules to remain within certain limits; it is for their mean kinetic energy to stay within certain bounds. For the water's temperature to change is for their behavior to exceed those limits. What we call "conditions" are patterns of *constant* change (i.e., of change that occurs in a uniform manner).

What we call "events" are *changes in the manner of change.* When the water's temperature increases, the *way* its constituent molecules move about changes. When the water's temperature stays the same, they still move about —they still undergo change—but the *way* they do so doesn't undergo change: those changes fall within certain limits. So conditions *can* be causes. But since conditions, far from being the absence of events, *consist* of them, this is in no way inconsistent with our contention that it is events, and events alone, that are causes.

2.2 All causes *ipso facto* spatiotemporal

There are causes only where there are changes, and changes happen in time. So universals, if they exist, are causally impotent. Many have concluded that, for this reason, they cannot be known and, therefore, that there is no legitimate reason to grant their existence. The idea is that there must be some kind of a causal connection between two things if one of them is to know of the other. For me to see the rock, it is necessary that light rays bouncing off of it affect certain sensory surfaces of mine. For me to know of Socrates' existence, I must be on the receiving end of testimony, the vehicle for which is always some causal process (noises, light rays bouncing off of ink marks), that can be traced back to some state of affairs involving Socrates. One obvious way to generalize this line of reasoning is to say that there is *never* awareness where there isn't a causal nexus.

For reasons given in Chapter 12 of this book (and given in greater detail Chapters 8 and 9 of my (2009) book *Analytic Philosophy* and in Chapters 6-9 of my (2007) book *Conceptual Atomism and the Computational Theory of Mind*), this conception of conception is erroneous; under no circumstances does a causal connection between two things *by itself* mediate the one's having an awareness of the other. There is always a non-causal component to awareness, even to sense perception, which, of all the many forms of awareness, is the one in which a causal connection between subject and object is most obvious. Further, for reasons given in Chapter 2, we do have good reasons to countenance the existence of universals, and no good reasons not to.

3.0 Causes as INUS-conditions

The things we describe as "causes" are really *partial* causes. Supposing it correct, the sentence "Smith's drunkenness is what caused the accident" must really be elliptical for "Smith's drunkenness is a *part* of what caused the accident." Had he been driving a car with new tires on an open country road, instead of a car with bald tires on a busy city street, Smith might not have crashed *even* if he'd been drunker than he actually was. So it wasn't Smith's drunkenness or any other one thing that caused the crash, and was instead a confluence of circumstances. In some contexts—for example, those of a legal or ethical or psychological nature—the *relevant* partial cause might well be Smith's drunkenness; and that seems to be what we mean when we say that it was *the* cause.[318]

The *complete* cause of anything is very hard to identify. It is a matter of philosophical and scientific debate whether, given an event *E* happening at *t*, anything other than the entire state of the universe just prior to *t* can be the *complete* cause of *E*. So for all intents and purposes, the expression "*the* cause of occurrence *E*" never refers to a *sufficient* condition for *E*.

"Even though Smith's being drunk wasn't *sufficient* for his crashing the car," it will be thought, "it was *necessary* for it. Had he *not* been drunk, he wouldn't have crashed." This is false. Smith's being drunk wasn't necessary for his crashing. If he had been sober, but had been driving too fast on an icy road, he might well have crashed.

J.L. Mackie (1917–1981) made it clear what is meant by "the cause of *E*," as this expression is typically used. Smith's being drunk wasn't a necessary condition for his crashing.[319] He could have crashed if sober. But his being drunk was a necessary condition for his crashing *under the circumstances*. Given that he wasn't driving a high-performance car on a country road during a clear day, and given that he *was* driving a low-performance car on a rainy night on a busy street, Smith's being drunk *was*, we may suppose, a necessary condition for his crashing. For it could be that, had he been sober, Smith would have narrowly avoided crashing, despite the rain, darkness, and faulty steering mechanism, and that it was only these things *plus* Smith's drunkenness that led to the crash. So, if it is to be correct, the statement "Smith's drunkenness was *the* cause of the crash" must be taken to mean that, *under the operative circumstances*, Smith's being drunk was necessary for his crashing.

There are just a few more points to make. First, given that Smith did crash the car, it follows trivially that *a* sufficient condition for his doing so was fulfilled. Second, whatever that sufficient condition was, it wasn't *necessary*. Suppose that what led to the crash was Smith's driving a low-performance vehicle on a dark, rainy night, while drunk. Even though, by hypothesis, that confluence of circumstance was sufficient for his crashing, it wasn't *necessary*. Smith might have crashed even if he'd been driving a high-performance vehicle in broad daylight while stone cold sober.

Thus, Smith's being drunk was an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition for his crashing. Smith's being drunk was thus an "INUS-condition" for his crashing. For this reason, Mackie says that if it is to express a true statement "*E* was *the* cause of event *E**" must be taken to mean that *E* is an INUS-condition for *E**. In a word, "the cause" of an event is an INUS-condition for that event.

4.0 Aristotle on causality

Aristotle (384–322 B.C.) distinguished between four different kinds of causality: efficient, final, material, and formal.[320]

> (AC)[321] *x* is the efficient cause of *y* if *x* makes *y* happen.
>
> *x* is the final cause of *y* if *y* is an action on the part of some creature and *x* is that creature's objective in performing *y*. Final causes are goals or objectives.
>
> *x* is the material cause of *y* if *x* is the material that *y* is made of.
>
> *x* is the formal cause of *y* if *x* is *y*'s form. (To expedite discussion, we'll henceforth omit the obvious qualifications relating to the fact that Aristotle didn't use English words.) Sometimes Aristotle uses the word "form" to refer to shape. But sometimes he uses it to refer to what he elsewhere refers to as "essences." Those properties that are essential to a given thing jointly constitute its essence; and a property is essential to a given thing if that thing could not lose that property without *for that very reason* ceasing to exist. The italicized phrase is intended to indicate that the dependence-relation in question is not of a *causal* nature, and is instead of a *logical* or *constitutive* nature. In other words, that relation is comparable to the dependence that triangles have on being three sided, not to the dependence that mammals have on air.

Setting aside a few trivialities, pretty much nothing Aristotle has to say about causality has any merit. Aristotle incoherently believes that objects *simpliciter*---as opposed to objects' having these or those properties---have essences. This belief, as we'll soon see, is an incoherent one, and along with the more general coherence of which it is but one instance, it vitiates Aristotle's views on causality.

It is essential to a thing's *being triangle-shaped* that it have three sides. But a given triangle-shaped object's continued existence needn't depend on its continuing to have three sides. Things change shape. It is essential to a thing's *being water* that it consist of H_2O molecules. But a given puddle's continuing to exist needn't depend on its consisting of such molecules.

Things undergo chemical changes without ceasing to exist.

Contrary to what Aristotle thought, dependence-relations don't hold between objects. They hold between *events* or between *statements,* depending on whether they're of a causal or a logical nature. To say that *x*'s having three sides is essential to *x*'s being a triangle is to say that "*x* is a triangle" entails that "*x* has three sides." To say that *x*'s consisting of H_2O molecules is essential to its being water is to say that "*x* consists of water" and "*x* doesn't consist of H_2O molecules" are incompatible statements. To say that treating children with kindness is essential to their becoming sane adults is to say that treating a child unkindly prevents him from becoming a sane adult.

So-called formal causes don't hold between events (or states of affairs) or statements; they hold between things. Thus, the concept of a formal cause is an incoherent one. So, for the same reason, is the concept of a material cause.

The concept of efficient causality clearly *is* a coherent one. There is such a thing as efficient causality. We'll see in the next section that it isn't the only *bona fide*form of causality. But it's the most fundamental one.

What about (so-called) final causes? If, as Aristotle seems to think, final causes are goals per se, he's wrong to describe them as causes. Goals per se do nothing. Goals are abstract objects and therefore do no causal work. What *does* do causal work is a given person's *having* a given goal. My *having* the goal that:

(*) I, JMK, get in shape;

obviously does causal work; it causes me to jog around the block for hours on end. But (*) per se does nothing. Like all other goals, (*) is a proposition. My *having* that goal is my intending to make that proposition come true. But my *having* a goal is a garden-variety efficient cause. In general, goals per se, being propositions, aren't causes at all. A given person's (or creature's) having a given goal *is* a cause. But it's a garden-variety efficient cause. So in saying there are such things as "final" causes, Aristotle is either making the trivial point that a person's having a goal can have effects *or* he is making the deeply confused point that goals per se can have effects.

4.1 The importance of the expression “in virtue of” in causal discourse

The expression “in virtue of” has an important place in philosophical discourse. (I will use “in virtue of” and “by virtue of” interchangeably.) Without this expression, or some equivalent, it would be very hard to say anything meaningful about causal explanation—or about non-causal explanation, for that matter. The meaning of this term is to be understood in terms of two parallel points, namely:

> (i) It is states of affairs (or events), not objects, that have effects;

and:

> (ii) it is propositions, not objects, that stand in dependence-relations.

For reasons that will become apparent, (i) and (ii) can be condensed into:

> (iii) dependence-relations hold between propositions, as opposed to objects.

A story will help us explain what (i)–(iii) mean and why they’re true. Despite being intellectually mediocre, Smith never loses an argument. He wins arguments, not by addressing points with counterpoints, but by bullying his opponent into submission. In this way, Smith recently won an argument with Jones, who is Smith’s superior intellectually. Thus:

> (A) even though Smith won his argument with Jones, he did so not *by virtue of* making good points, but *by virtue of* bullying Jones.

What (A) means is that the cause of Smith’s winning was *his being a bully*, and not *his being smart.* (Cause: Smith’s being a bully. Effect: Smith’s winning. *Smith* wasn’t the cause of anything.) Thus, the occurrence of “by virtue of” in (A) is a way of encrypting a statement about states of affairs—of making them seem about objects, when in fact they are about states of affairs.

Even if Smith *did* make good points in the course of the argument, it doesn't follow that he won *by virtue of* making good points. It could be that his making good points had nothing to do with his winning and that it was his bullyism that was responsible for it. In that case:

> (A*) even though Smith won his argument with Jones, and even though he made good points in the course of that argument, Smith won *by virtue of* his being a bully, and not *by virtue of* his making good points.

Importantly, Smith could have won *both* by virtue of being a bully *and* by virtue of making good points. In other words, his bullyism and his making good points could have been co-factors.

Let us now discuss (ii). Suppose that there are only three triangular objects in the world, *x* being one of them, and also that these three objects are the only green objects in existence. In that case, although "every triangle is green" is true, "*x* is green *by virtue of* its being triangle" is false. The right statement is:

> (B) "*x* is a triangle by virtue of its having three co-planar sides, each of them straight, such that any two of them intersect and such that not all three of them intersect."

What (B) means is:

> (C) The proposition that *x* is a triangle entails, and is entailed by, the proposition that *x* has three co-planar sides, each of them straight, such that any two of them intersect and such that not all of three of them intersect.

In what follows, I will use the term "fact" to mean "spatiotemporal state of affairs." So a "fact," as I'll be using this word, is a distribution of mass-energy.[322]

Suppose that, for reasons of physical law, nothing could be triangular without being green, and nothing could be green without being triangular. In that case, given some green/triangular object *x*, the fact in virtue of which *x* is green—in other words, the distribution of mass-energy in virtue of which *x*

was green—would be spatially and causally indistinguishable from the fact in virtue of which *x* was triangular. Nonetheless, it wouldn't be in virtue of its being a closed three-sided (etc.) figure that *x* was green—it would be in virtue of its having certain reflective properties. And it wouldn't be in virtue of its having those reflective properties that *x* was triangular—it would be in virtue of its being a three-sided, straight-edged (etc.) figure. This shows that the meaning of (B) is (C). When occurring in statements affirming logical (as opposed to causal) dependencies, "in virtue of" denotes a relationship holding, not between things *or* even facts, but between propositions.

Given that the occurrence in (A) of "by virtue of" does denote a relationship holding between states of affairs, what are we to say? One response, of course, is to say that it's ambiguous. But, in addition to appearing very wrong, this response is unnecessary. (A) doesn't express an *entailment.* But it does express *a* dependence-relation holding between statements. The truth of *Smith won the argument* is, in this context, dependent on the truth of *Smith bullied his opponent.* In general, "in virtue of," when occurring in causal contexts, can be interpreted as denoting a form of interpropositional dependence; and we may therefore condense (i) and (ii) into:

> (iii) Dependence-relations hold between propositions, as opposed to things.

4.2 Functionalism, content-externalism

According to a doctrine known as "functionalism," it is *wholly* in virtue of what its causes and effects are that a given brain-state (or physical entity of any kind) falls into a given mental category. Anything that has certain causes and certain effects is *ipso facto* a belief that snow is white or a desire for a world peace or a vitriolic dislike of a certain piece of music.[323] Functionalism is discussed at length in Chapter 22. But a few words on it here are appropriate.

Functionalism has no plausibility when it comes to sensations or, more generally, anything that has ***phenomenal content.*** (For a mental state to have "phenomenal content," or "phenomenology," is for there to be a certain kind of consciousness such that, having that form of consciousness is necessary and sufficient for being in the state. So feelings all have phenomenal content; so do all perceptions. But enduring mental structures---judgments, attitudes, values---lack phenomenology. Of course, they give rise to conscious events that have phenomenal content. But they themselves lack it, for there is no specific kind of consciousness such that having it is necessary and sufficient for believing that 1 + 1 = 2 or being suspicious of one's neighbor's intentions.) But functionalism is widely taken to be correct when it comes to so-called "dispositional" mental entities such as beliefs and objectives. (I put the word "dispositional" in quotes, because, for reasons that I'll now give, I reject this way of characterizing the mental entities in question.)

4.2.1 Why, even thus restricted, functionalism is indefensible

Thus restricted, functionalism may initially sound plausible. But it has some grisly problems that a story will help to expose. While standing right in front of you, Jones, a demented acquaintance of yours, puts what is clearly arsenic in a glass of water. He then tries to make you drink the water. Jones is very strong and is also a trained martial artist; so while holding the glass in one hand, without spilling any of it, he puts you in a joint-lock and tries to pour the contents of said glass down your throat. You shut your mouth, and you twist and squirm. Miraculously, you break free and run out the door.

What caused all of this behavior? A good part of it was your belief that Smith was trying to make you drink arsenic. Another part of it was your belief that, were Smith to succeed in making you drink arsenic, you'd suffer some adverse consequence. Yet another part was your wish to not die. And, of course, other ***propositional attitudes*** of yours were determinative of your behavior. Supposing that *x* is the part of your brain that comprises the various physical structures and events that mediate these different propositional attitudes, it isn't just *x* that makes you squirm and fight, and so on. Rather, it's *x's being a belief that Jones is about to make you drink arsenic.*

But functionalism cannot accommodate this obvious fact. According to functionalism, it is in virtue of what *x* causes and what causes it that *x* is a belief that Jones is about to do that. The causal relations fix the content, not the other way around. This means that, if functionalism is right, *x*'s being a belief that Jones is about to make you drink arsenic does *nothing.* In general, given any entity O that mediates mental state *M*, *O*'s being an instance of *M* is causally inert. But this means that, if functionalism is correct, then the mental does nothing; it has no causal powers at all. Of course, functionalism is consistent with, and in fact depends on, the presumption that the brain-states that mediate mental states have causal powers. But since functionalism is inconsistent with the fact that *in virtue of being mental states,* brain-states are powerless, functionalism amounts to the denial that the mental (*qua* mental[324]) has any causal powers at all. Given the fact that anything spatiotemporal *ipso facto* has causal powers, the reason being that occupancy of a time or a place *is* a form of causality, it follows that functionalism denies the very existence of minds.

4.2.2 Content-externalism: another popular but indefensible doctrine

According to a doctrine known as "***content-externalism,***" two people, Smith and Twin-Smith, who are exactly the same, leaving aside the remote causes of their current conditions, can have thoughts with different contents. So, because Smith's current visual perception was caused by light that bounced off of Mary, whereas Twin-Smith's current visual perception was caused by light that bounced off of Twin-Mary, otherwise identical Smith and Twin-Smith are now having thoughts with different truth-conditions and, therefore, different contents. (Smith is thinking *that person*—viz., Mary—*is tall.* Twin-Smith is thinking *that person*—viz., Twin-Mary—is tall.)

To appreciate just what the content-externalist says, we must distinguish causal from constitutive necessity. Obviously people have different thoughts. And obviously differences in people's personal pasts *create* many of the cognitive divergences between them. But the content-externalist is saying that, given two people who, like Smith and Twin-Smith, are *now* atom-for-atom duplicates of each other, and whose respective pasts haven't so much as made a dent in the one person that they haven't made in the other, those two people *will think different* things. They'll have thoughts with different contents. One of them will think about Mary (gold, water, etc.), while the other thinks about Twin-Mary (fool's Gold, faux-water, etc.)

Content-externalism is incoherent. If Smith and Twin-Smith were to switch places, but everything else were held constant, neither would behave or think differently from how he would otherwise do so. Thus, *to the extent* that Smith's mental content is fixed by factors that aren't operative in Twin-Smith's case, Smith's thinking isn't at all different from Twin-Smith's. Any differences between Smith's and Twin-Smith's thoughts have some basis other than the differences in their distal causes. So to the extent that mental content is to be understood along content-externalist lines, it is without any causal powers at all, and therefore doesn't exist.

As for the content-externalist's contention that Smith is thinking that *Mary* is tall, whereas Twin-Smith is thinking *Twin*-Mary is tall, that is a non-starter. What Smith's eyes tell him isn't that *Mary* is tall. They tell him that *a* person who has certain features is tall. Given only what his eyes are telling him, Smith doesn't know whether he's seeing Mary or some Mary look-alike. To be sure, he may well know that he's looking at Mary. But that knowledge

won't be based strictly on the content of his visual perception. It will be based also on background knowledge that he brings to the present context.

Content-externalism embodies a failure to appreciate the concept meant by the phrase "in virtue of." True-Smith and Twin-Smith aren't doing the same things. Smith is in one place; Twin-Smith is in another. Smith is looking at one person; Twin-Smith is looking at another. But it isn't *in virtue of* any differences in their respective mental contents that Smith and Twin-Smith diverge in these respects. Since, therefore, no differences between Smith and Twin-Smith, or by obvious extensions of this argument, between any two people, are to be understood along content-externalist lines, mental content is causally impotent, and therefore non-existent, to the extent content-externalism is the right theory about it.

Tyler Burge (1946–), a leading exponent of content-externalism, deals with this by arguing that the notion of causality should be revised, so as to accommodate content-externalism. Were two qualitatively identical objects *x* and *y* to switch places but everything were held constant, this switch would be experimentally undetectable. Burge acknowledges this but says that, nonetheless, *x* and *y* have different causal properties. By thus reanalyzing the concept of causality, he safeguards content-externalism from the objection to it just considered.

In addition to being brazenly *ad hoc*, this is incoherent, since it says, quite straightforwardly, that there can be distinctions without corresponding differences.

Also, the relevant data is easily modeled without stooping to such lows. To see why, see Chapters 6, 8, 9, and 18 of my (2009) book *Analytic Philosophy*, as well as my (2010) article "Intensionality and Modality," published in the Journal of Pragmatics, and, finally, pages 207–507 of my (2007) book *Conceptual Atomism and the Computational Theory of Mind.*

4.3 Program-causality[325]

We obviously know of many cases of one event's causing another. We are inclined to think that, in each such case, the one event is the *efficient* cause of the other. Working jointly, Frank Jackson (1943-) and Philip Pettit (1945–) have shown this to be false. Given almost any known instance of one event's causing another to occur, the one event is the *program-cause,* not the *efficient* cause, of the other.

What does the term “program cause” mean? Some scenarios will help us answer this question. There is a kettle on the stove. You turn on the gas. The resulting flames underneath the kettle cause it to heat up and thus cause the water to boil. Let F be the flame. We’re inclined to think that *F* caused the kettle to heat up.

> (i) *F* (or, in any case, *F*’s having a certain temperature) caused the kettle to heat up.

But only a tiny part of *F*’s surface was in contact with the underside of the kettle. Had the events at those contact points been present, the effect on the kettle would have been the same, even if the other flame-constitutive events were absent.

Another example: Statements such as:

> (ii) high-inflation led to crime, which, coupled with xenophobia on the part of the populace, gave the government a powerful incentive to scapegoat a foreign nation and go to war with them.

But causes are much more specific than (ii) suggests. *Crime* doesn’t do anything, and neither does *xenophobia*. Things are caused by specific people, performing specific criminal or xenophobic acts.

One last example:

> (iii) The impact of the rock on the window caused the window to shatter.

This case is similar to the first. Only a small part of the surface of the rock came into contact with the window—and, in fact, those parts of the rock’s surface came into contact with specific parts of one of the window’s two surfaces. Had the events at those contact points happened, the effect on the window would have been the same, even if no rock had struck the window.

But if we say that (i)–(iii) are just wrong, our grounds being that a “cause” of an event *E* can only be an event that is directly responsible for *E*’s occurrence, then just about every single causal statement anyone makes will be wrong. But statements like (i)–(iii) obviously convey valuable

information. So rather than say that they're all just wrong, we should look for an interpretation of the word "cause" that ***validates*** those statements.[326]

Jackson and Petit have done just that. The form of causality mentioned in (i)–(iii) is *program-causality*. *x* is a program-cause of *y* if *x* consists of events that, in their turn, are directly responsible for *y*. A country's being poor consists of various specific states of affairs—various specific people being without funds in various specific situations. The specific situations that *constitute* a condition of wide-spread poverty will directly cause crimes. And so on. *F* heats up the pan in the sense that among the events constitutive of *F* are those that heat up the pan; and it is in the same sense *mutatis mutandis* that the impact of the rock shattered the window. Thus, Jackson and Pettit have successfully validated ordinary causal statements.

They make other important points. One would think that high-resolution causal information is better than low-resolution causal information. But this isn't always so. What caused the house to burn down?

> (SA)[327] Arson on Smith's part was the cause. He threw several Molotov cocktails at the house and up in flames it went.

SA obviously doesn't give us the micro-story. For argument's sake suppose that some long sentence MS does give us the micro-story. MS will be worthless in terms of giving us information about Smith's character, and it will be equally worthless in other, comparable respects. There are many principled regularities of which this particular incident is an instance, and we blind ourselves to all but a few of them if we reject any story that isn't the micro-story. So the macro-story—or, rather, the various different macro-stories—aren't inferior versions of the micro-story; they aren't precursors to it that we'll chuck when we have it. They expose regularities, structures, and counterfactual truths omitted by the micro-story. This shows that biology isn't just bad physics, that psychology isn't just bad physiology—it shows that different disciplines are dealing with different program-causes, and therefore distinct and very possibly *sui generis* nomic connections—that happen at least sometimes to be instantiated by the same efficient (micro) causes.

Predictions are almost always made on the basis of a knowledge of program-causes. To predict what's going to happen, one must know of *structures* that can be depended on to yield certain outputs, given certain

inputs. So far as one lacks such knowledge, one must make predictions on an ad hoc basis, and there are two reasons why this is a problem. First, sub-structural facts tend to be small—literally small. (Something is a structure only to the extent that it's a cohesive aggregate of events. Something fails to be a structure only to the extent that it's an isolated event. Thus, lone events tend to be smaller than structures by orders of magnitude.[328]) And when they can be known, the process of learning about them tends to disrupt them in prediction-foiling ways. (See Chapter 14.)

Also, structures comprise back-up mechanisms that ensure a certain outcome even if, because of some hidden variable, some rogue, structure-unfriendly event makes it through the gauntlet. If a salesperson working for megacorporation *X* unexpectedly quits, another takes his place. If one police officer doesn't report for duty, another steps into the breach. It's of the essence of structures that a certain amount of causal redundancy be incorporated into them. Where there isn't such redundancy, predictions are easily foiled, since there is nothing to absorb the ripple-effects created by some rogue occurrence.

Given the dependence of predictive knowledge on structures and, therefore, on program-causes, it follows that knowledge of explanations is comparably dependent. The reason is that explanation is reverse-prediction. One predicts *X* when one identifies current conditions that, given known mechanisms, make *X*'s occurrence likely. One explains *X* when one identifies past conditions that, given known mechanisms, resulted in *X*'s occurrence. The concept of a program-cause, which was, so far as I know, first clearly analyzed by Pettit and Jackson, is therefore central to the concepts of prediction and explanation and, thus, to the theory of knowledge.

Chapter 15 Hume’s First Analysis of Causality

1.0 Hume's counterintuitive analysis of causality supposedly *de rigueur*, given his acceptance of empiricism

According to David Hume, nothing makes anything happen. There are no forces or structures underlying the various events that occur. There are just the events.

Why does Hume think this? Because he accepts a very strict form of ***empiricism,*** this being the doctrine that what is known is either (i) directly observed to be the case, or (ii) is inferred from what is directly observed in accordance with inference-rules of whose legitimacy direct observation has apprised one.

Do we observe connections between events, Hume asks? No—we only observe the events. Thus, Hume believes, so far as you have empirical grounds for saying that the collision of billiard balls B_1 and B_2 caused previously stationary B_2 to start moving, it's because you know the following three conditions to be fulfilled (let E_1 be the collision and let E_2 be the subsequent displacement of B_2):

> (HC)[329] (1) E_1 immediately preceded E_2; (2) E_1 and E_2 are adjacent; and (3) any event similar to E_1 immediately precedes, and is adjacent with, some event similar to E_2.

Conditions (1) and (2) explained: If there were a spatial or temporal gap between the two events, we'd be unjustified in believing the first to be a *direct* cause of the second. Hence conditions (1) and (2). Condition (2) explained: If events similar to E_1 weren't typically adjacent with, and immediately followed by, events similar to E_2, we'd have no reason to think that E_1's occurring had anything to do with E_2's occurring.

Hume thus proposes the following analysis of causation:

> (HC_1)[330] E_1 causes E_2 to occur if E_1 immediately precedes and is adjacent with E_2, and any event similar to E_1 immediately precedes and is adjacent with some event similar to E_2.[331]

2.0 Problem #1 with Hume's analysis: It doesn't distinguish between actual causation and pseudo-causation

This analysis is untenable. A story will help us identify one of the reasons why. Your dog Fido is an infallible earthquake detector. Whenever an earthquake is about to strike, he senses it, and barks hysterically and runs around in circles. One day, Fido starts barking hysterically and running around in circles. Sure enough, an earthquake strikes right then and there.

Hume's analysis entails that Fido's antics *caused* the earthquake to strike. Fido's antics immediately precede the earthquake. (So condition (1) is satisfied.) Fido's antics are adjacent with it. (So condition (2) is satisfied.) And such antics on Fido's part are *always* adjacent to an immediately ensuing earthquake. (So condition (3) is satisfied.) Given that Fido's antics obviously didn't cause the earthquake, Hume's analysis is wrong.

The real cause of the earthquake presumably created some disturbance that Fido detected, that being why he engaged in his antics. Thus, some one thing was responsible both for Fido's antics and for the earthquake. The two effects fulfilled the conditions described by Hume. For this reason, the sequence they formed mimicked a *bona fide*causal connection.

When some one thing E_1 has two effects E_2 and E_3 that *appear* to be related as cause and effect, respectively, we say that the sequence consisting of E_2 and E_3 is a "pseudo-causal process" or, alternately, a case of "epiphenomenal causation." Hume's analysis cannot distinguish epiphenomenal causation from actual causation.

3.0 Problem #2: Hume's analysis entails that events can't be space-time neighbors without one of them causing the other to occur

If E_1 and E_2 are any two spatiotemporally adjacent events, it's easy to find categories K_1 and K_2 such that E_1 and E_2 belong to K_1 and K_2, respectively, and such that any instance of K_1 immediately precedes, and is adjacent with, some instance of K_2. Though he doesn't know it, Bob is due for a heart attack. Right before the heart attack, Sally gently hugs him. Given the almost certainly correct supposition that nobody who is *exactly* Sally's height has ever hugged anyone who is *exactly* Bob's height, it follows that any hug involving people with heights m (Sally's height) and *n* (Bob's height) immediately precedes and is adjacent with a case of somebody of height n having a heart attack. (So if K_1 is the category of hugs involving people with heights *m* and *n*, and K_2 is the category of people with heights *n* having heart attacks, any event E_1 that belongs to K_1 immediately precedes and is adjacent with some event E_2 that is a member of K_2.) So Hume's analysis has the false consequence that the hug caused the heart attack. Given any two spatiotemporally adjacent events *x* and *y*, a similar argument shows that, if Hume's analysis is right, *x* caused *y*. Since not all such pairs of events are related as cause and effect, Hume's analysis is false.

3.1 Problem #2 (continued): No way to fix Hume's analysis without making it viciously circular

In response to this it will be said that:

> The categories in question must be relevant ones. K_1 and K_2 won't do. They're the wrong ones. All you've shown is that Hume's analysis must be modified. Instead of:
>
> > (HC$_1$) E_1 causes E_2 to occur if E_1 immediately precedes and is adjacent with E_2, and any event similar to E_1 immediately precedes and is adjacent with some event similar to E_2,

Hume should have said:

> (HCI#) E_1 causes E_2 to occur if E_1 immediately precedes and is adjacent with E_2, and any event *relevantly* similar to E_1 immediately precedes and is adjacent with some event *relevantly* similar to E_2.

This proposal doesn't work. Given any two spatiotemporally adjacent events, each of those two events will belong to many different categories. Hume's analysis is powerless to tell us whether the one caused the other *until* we select the relevant categories. *But we can't select the relevant categories until we have the very knowledge that Hume's analysis is supposed to give us.* It is only because we already know that the hug isn't what caused the heart attack that we know that K_1 and K_2 aren't relevant categories.

A related point is that HC1# is viciously circular. Supposing that E_1 directly causes E_2, for K_1 and K_2 to be the "relevant" categories is for it to be the case that it's *because* E_1 and E_2 belong to them that those events are related as cause and effect. Thus, HCI# collapses into the viciously circular statement that:

(HC1+) E_1 causes E_2 to occur if E_1 immediately precedes and is adjacent with E_2, and anything like E_1 causes something like E_2 to occur.

4.0 Problem #3: Hume's analysis falsely assumes that spatiotemporal position is non-relative

According to Hume, *x* is a cause of *y* only if *x* and *y* are space-time neighbors. Thus, Hume analyzes the concept of causality in terms of the concepts of space and time. Were it to turn out that the concepts of space and time are *themselves* to be understood in terms of the concept of causality, Hume's analyses would be viciously circular. He'd be analyzing causality in terms of itself and thus not analyzing it at all.

In fact, the concepts of space and time *are* to be understood in terms of the concept of causality, and Hume's analysis *is* viciously circular for this reason. According to modern physics, E_1 precedes E_2 iff it's possible for a signal beginning with E_1 to be sent to E_2. E_1 and E_2 are simultaneous iff it isn't possible for a signal beginning with either to be sent to the other. Order in time is thus to be understood in causal terms.[332]

So is order in space. Events happen in different places if, and only if, something can affect the one that cannot affect the other. If *x* and *y* are in coincident places, nothing can affect the one without affecting the other. If *x* and *y* are not in coincident places, something can affect the one without affecting the other. Thus, *x* and *y* are in different places exactly if something affecting the one may fail to affect the other. So if one says that *x* and *y* occupy distinct but adjacent places, or that they occupy some one place, one is *ipso facto* making a causal statement. Thus, Hume's analysis of causality is circular, given that it analyzes ‹*x* causes *y* to occur› in terms of ‹*x* is spatiotemporally adjacent with *y*.›

4.1 Problem #3 (continued)

Here is a similar line of thought. It would be meaningless to say that object *x* moved but that its causal liaisons to other objects didn't change for that very reason. If I approach you, my causal relations to you change, and so do your causal relations to me. After those changes are set aside, there is nothing left: there is no "pure" change in position underlying all the causal changes.

Also, the concept of spatial occupancy is a causal concept. What is it for me to be occupying this place? It is for it to be harder than it would otherwise be for you to enter this place. It makes no sense to say: "*x* is occupying spatial region *R*, but *x*'s occupancy of *R* has no effect on what things other than *x* must do to occupy *R*." To occupy a place *is* to make it harder, other things equal, than it would otherwise be for other things to enter that place.

Hume analyzes causality in terms of the concept of spatial occupancy. E_1 causes E_2 only if they occupy adjacent, or coincident, spaces. But since E_1's (or E_2's) occupancy of a place is *itself* to be understood in causal terms, Hume's analysis analyzes causality in terms of itself and therefore doesn't analyze it at all.

A corollary of the fact that spatiotemporal relations are causal relations is that, contrary to what Hume alleges, we *do* see causal processes and, moreover, we see them *as* causal processes. We see nothing instantaneous. Everything we see persists, if only for a short while. Anything that persists *is*, for that very reason, a causal process; and anything that is seen *as* persisting, which is everything that is seen, is, for that very reason, seen *as* a causal process.

4.2 Two conceptions of spatiotemporal position

There are two views as to what space and time are. According to the one view, space and time are not empty containers. Spatiotemporal positions are relations between events, and there thus couldn't be a space-time empty of events. If this view is right, it's meaningless to say some object *just* approaches some other. For *x* to precede *y* isn't for *x*'s *position* to bear a certain relation to *y*'s position. Given *any* spatiotemporal relation *R*, for *x* to bear *R* to *y* is not for *x*'s position to bear *R* to *y*'s position. There are no event-independent positions. Relations between positions are mediated by relations between events, not *vice versa.* It thus makes no sense to suppose that there could be bare spatiotemporal interrelations or, therefore, bare changes in such interrelations.[333]

According to the other view, space and time exist independently of events. A consequence is that spatiotemporal positions exist independently of the events that might (or might not) occupy them. Spatiotemporal relations holding among events are to be understood in terms of spatiotemporal relations holding among positions in space-time. This means that there are absolute positions in space-time. If this view is right, it *does* make sense to say that one event *just* precedes some other event, or that one event *just* happens to be equidistant between two other events, and so on.[334]

Bearing these points in mind, consider the statement that:

(IP)[335] *x* immediately precedes and is adjacent with *y*,

and suppose IP to be true. If the first view is the right one, there are no event-independent positions. In that case, *x*'s being adjacent with *y* is a relation that holds *directly* between *x* and *y*, and to say that some position bears a certain relation to another position is to say that some *event* bears a certain relation to some other event.

If the second view is the right one, the fact described by IP consists in *x*'s occupying some position that bears a certain relation to a position that is, in turn, occupied by *y*. In other words, it is only because the *positions* they occupy are adjacent that *x* and *y* are adjacent. Thus, the second view presupposes that there are event-independent, and thus absolute, spatiotemporal positions.

Since Hume's analysis presupposes the correctness of the second view,

Hume’s analysis is refuted by the very empirical and philosophical arguments that show the concept of absolute position to be an untenable one. (Further argumentation to the effect that the relational conception of space-time is the right one will be provided in this chapter, in Sections 4.2.2-5.0, and in Chapter 20, Section 8.0.)

4.2.1 Continuity identical with causality

Ball *A* crashes into previously stationary ball *B*, whereupon *B* lurches into motion. This is Hume's paradigm of an instance of causality. This change in *B*'s condition appears discontinuous. It would not seem that way were we to have a more high-resolution understanding of the exchange of energy that is actually constitutive of the collision.

In general, the more accurate our knowledge of what is involved in any apparently discontinuous change of state, the more we find that no discontinuities were involved. Given information about the aforementioned collision that is slightly more high-resolution than the sort of information about collisions that was available to Hume, it proves to be no more discontinuous than was *A*'s pre-collision state of motion. I now hope to show that Hume was very much led astray, so far as his understanding of causality was concerned, by the technological limitations of his time.

4.2.2 Continuity identical with causality (continued)

Hume's paradigms led him to see *discontinuity* as a *sine qua non* for causality. Let event *E* be the aforementioned collision, and let *E** be the subsequent change in *B*'s state of motion. For Hume, *E*'s being the cause of *E** doesn't involve *E**'s being the continuation of some process initiated by *E*. There is no torch that *A* is passing to *B*.

"How," it will be asked, "can you say that, for Hume, *discontinuity* is essential to causality? For the very opposite is clearly true. In Hume's view, cause and effect must be spatiotemporally adjacent. So for Hume, the essence of causality *is* continuity."

Discontinuity takes many forms. Like Hume, the objector is focusing on only one of the many relevant forms of continuity. For Hume, *E*'s causing *E** to occur involves *B*'s lurching into motion *spontaneously*. True---this sudden lurch into motion is spatiotemporally adjacent with the place where *A* stops moving. But the two events are otherwise disconnected. Also, *A*'s suddenly ceasing to move is, on Hume's view, just as spontaneous as *B*'s starting to move. So, on Hume's view, we have two discontinuities: *B*'s starting to move and *A*'s ceasing to move. So, for Hume, a causal nexus is a regularity whose instances are spatiotemporally adjacent *discontinuities.*

Also, Hume's conception of what it is for two events to be spatiotemporally adjacent is not a coherent one, a corollary being that, if Hume is right, *E*'s being the cause of *E** doesn't in any significant way require that one of those events be continuous with the first. *x* and *y* are two objects that are relatively near each other. *x* and *y* are spatiotemporally remote from *x** and *y** but are otherwise just like them. Moreover, *x** and *y** are spatiotemporally interrelated in exactly the same way as *x* and *y*: so *x**'s movements "track" *x*'s (so, for example, if *x* approaches *y*, *x** approaches *y** by the same amount), and *y**'s movements track *y*'s. But one day this changes. On that day, during interval *t*, *x* moves one foot closer to *y* than it was before *t*. during that same interval, *x** moves one foot closer to *y** than *it* originally was.[336] Otherwise, *x*'s relation to *y* is, with one important qualification, just like *x**'s relation to *y**.

But there is a difference. Because *x* is now one foot closer to *y* than it was previously, *x*'s causal relations to *y* have changed; but, despite being one foot closer to *y* than it used to be, *x**'s causal relations to *y** haven't changed at all. The amount of time that *ceteris paribus*[337] it takes light to travel from *x* to *y*

has decreased, as has the amount of energy[338] that *ceteris paribus* is needed to transport a given amount of grain from *x* to *y* has decreased; and *x*'s degree of gravitational attraction to *y* has *increased*. But the amount of time it takes light to travel from *x** to *y** has *not* decreased and neither has the amount of energy required to transport a given amount of grain from *x** to *y**; and *x**'s gravitational attraction to *y** hasn't *in*creased at all. So there is *nothing* to indicate that *x** has gotten one foot closer to *y** than it was before: the amount of time it would take *x** to travel to *y** hasn't changed; since *x**'s optical relationship to *y** can't have changed, the way *x** appears to somebody on *y** hasn't changed; there would be no reason of any kind---optical, kinematic, dynamic---to regard *x** as having undergone a change in position relative to *y**.

Spatiotemporal changes that aren't marked by causal changes are surds. But if spatiotemporal relations were more fundamental than causal relations, they would not be surds: bare (causally antiseptic) changes in relative position would be perfectly possible. Since they're not, it makes no sense to speak of changes in relative spatiotemporal position except as supervenient on changes in relative causal relations. Statements about spatiotemporal relations, we must conclude, are very generic, very low-resolution statements about causal relations.

So, in saying that *E* and *E** are *just* adjacent to each other, Hume is speaking incoherently. It makes no sense to speak of *mere* adjacency. Adjacency is a causal concept. The way in which *E* and *E** must be adjacent if Hume's analysis is to be free of vicious circularity must have no causal meaning: it must not be different, so far as *E*'s causal relation to *E** is concerned, from the relationship that *E* bears to some event that happened eons earlier, in some remote galaxy. But this eviscerates Hume's analysis, since that analysis obviously fails unless *E*'s relationship to *E** is relevantly different from its relationship to some long gone event in a distant galaxy. Since the relevant difference lies in *E*'s *causal* relations to *E**, Hume's analysis is doomed to vicious circularity. So, since Hume's concept of adjacency is hollow, his insistence that cause and effect be spatiotemporally adjacent is without any real meaning and doesn't undermine our contention that, for Hume, the occurrence of various *discontinuities* (of a certain kind) are a *sine qua non* for one event's causing another.

What Hume thought is the antithesis of the truth. The essence of causality is continuity. So far as Hume's belief to the contrary had any support, it lay in

the limited state of microphysical knowledge in his day.

4.2.3 Continuity with causality (continued)

Newton rightly said that the concept of "action at a distance" was "abhorrent" to anyone of any intelligence. These days, the conventional wisdom would be to say (i) that Newton's suspicion that causal influences must be propagated by intervening media was vindicated by later developments in physics but (ii) that there is nothing *absurd* in the supposition that there should be action at a distance.

I agree with Newton. The sentence "there could be action at a distance" is tautologously false. To say that *A* is a given distance from *B is* to say that it takes a certain amount of time for events occurring on (or next to) *A* to have an effect on *B*. To say that an event on *A* "instantaneously" affects *B is* to say that *A* and *B* are adjacent. So it is nonsense to say that, even though *A* and *B* are remote from each other, events on *A* could instantaneously affect *B*.

4.2.4 Continuity identical with causality (continued)

Suppose that object *A* disappears in place *P* and, at that instant, some perfect duplicate of *A*, which we'll call "*B*," appears in some place *P** that is remote from *P*. What grounds have we for saying that *B* is identical with *A*?

"What if we find out that *A*'s disappearance is what caused *B* to appear, seemingly *ex nihilo*, in *P**?"

But what conceivable grounds are there for saying that *A* 's disappearance *caused* B to appear? There was no lapse in time between those two events and thus no possibility of anything being transmitted from *P* to *P**.

"What if we find out that these two events---*A*'s disappearance and *B*'s sudden appearance elsewhere---is an instance of a general regularity? Wouldn't that be evidence of some causal connection?"

It would be evidence of epiphenomenalism causation, not of causation proper. It would be evidence that, underlying both *A*'s disappearance and *B*'s appearance, was some structure that, by way of continuous series of events, led to both.

"But, assuming, if only for argument's sake, that we've definitively established that there is no such structure and *a fortiori* that there are no such series, what we are to say then?" There are two ways to answer this. The first is to say: "We're dealing with a brute fact." The second is to say: "*A* and *B* *are* adjacent; and in stipulating that *P* and *P** are disjoint places, we were prejudicially ruling out a viable way of explaining the concomitance in question: we must rethink what it is to say that two events are spatiotemporally adjacent."

Strange as the second response may seem, it is, I believe, *de rigueur*, so long as we refuse to see the concomitance in question as an unexplainable, brute fact and so long as we know there to be series of events mediating between *A*'s disappearance and *B*'s appearance. We must remember what we said a moment ago about the concept of action at a distance. Newton believed this concept to be incoherent. Newton, we suggested, was right. One event's being adjacent with some other event is *identical* with the one's being able to have *immediate* effects on the other.

Assuming the truth of the assumptions within which we're currently working, I would propose that adjacency is a relative concept. Two events, *x* and *y*, may be adjacent *relative* to certain forms of causal influence and non-adjacent relative to other forms. It makes no sense, as we saw, to say that, by

virtue of changing its position relative to *B*, *A*'s causal relation to *B* failed to change. And the reason it makes no sense is that changes in position *are* causal changes. There are different ways in which one event can affect another. Let *E* and *E** be two different forms of energy, and let *x* and *y* be two events such that *E* takes a finite amount of time to travel from *x* to *y* and such that *E** takes no time at all to travel from *x* to *y*. Nothing can travel faster than light. (So *E** cannot travel faster than light.) Light takes a non-null amount of time to travel along *any* path whose length is greater than zero. So if it was by transmitting *E** that *A*'s disappearance brought about *B*'s appearance, then those two events *were* adjacent. That is, they were adjacent *relative* to *E**, even though they were also non-adjacent *relative* to *E*. In general, if we are forced to operate with the framework within which our adversary insists that we operate, the right way to model this data is to say: "Relative position is *itself* a relative notion. Given two events *x* and *y*, they may be adjacent *relative* to certain forms of energy and non-adjacent relative to others."

This model is a coherent one. It *seems* incoherent only if one assumes that there is event-independent spatiotemporal position. But judgments as to spatiotemporal position cease to have meaning the moment they cease to do anything more than register causal relations between the events in question. Let events *x* and *y* be such that:

> (i) *x*'s distance from *y* is non-null, given the amount of time that it takes causal process *K* to start with the one event and end with the second,

and such that:

> (ii) *x*'s distance from *y* is null, given the amount of time that it takes causal process *K** to start with the one event and end with the second.

Given (i) and (ii), we'd have to say that, relative to processes of the same kind as *K*, *x* is not adjacent with *y*, but that, relative to processes of the same kind as *K**, *x* and *y* are adjacent. It is *logically* possible that (i) and (ii) should both be true. And, by *doubly* relativizing spatiotemporal position in the way just described, we can hold onto the fecund principle that nothing travels

faster than light.

5.0 Problem #4: Hume's analysis isn't strictly empirical

We don't see light; we see illuminated surfaces. Similarly, we don't see space; we see spatially separated objects. Look around you—at your desk, at the painting hanging on the wall, on the door on the opposite side of the room, etc. These things are spatially separated from one another. And there is no doubt that you know this *through* observation. But do you observe *the space* that separates any two of these items?

You do not. You observe the items. On the basis of their appearances, along with background knowledge (e.g., the intensity, duration, and nature of the effort involved in your walking from your desk to the painting), you *infer* that the desk is separated by (say) a distance of approximately 10 feet from the painting.

The sense in which one perceives space and time is not comparable to the sense in which one perceives the objects in them. One perceives events. Those events all have various properties. Among them are properties that spatiotemporally order them. You *abstract* the concepts of space and time from your knowledge of these properties (or property-instances, rather). But you no more perceive space or time than you perceive propertyless substances. In fact, the concepts of an empty space and an empty time belong in the same category as the concept of such a substrate, and all three concepts belong in the same category as the concept of a square circle. Compare the statement:

> (1) A day has just passed since I last experienced gastrointestinal pain, and during that day not a single event occurred in the entire universe,

with:

> (2) I *just* experienced gastrointestinal pain.

In a world in which (1) was true literally nothing would have happened to affect the speaker's relation to his last episode of gastrointestinal pain. There would be nothing in the condition of any constituent, no matter how big or how small, that would distinguish a world where (1) was true from one where (2) was true.

To speak meaningfully of there being a *passage* of time, there must be events. Otherwise I'd be perfectly in my rights to say that between the time at which I began to type this very sentence and the time at which I stopped typing the previous one there was an interval of a day, or of a billion years, when in actuality the interval was less than one second.

Statements about the "passage of time" are very abstract statements about events; they're statements about the events one has experienced that register such generic properties of events that one can assume the events in one's interlocutor's experience to have relevantly similar properties.

It is thus through *conception*, not *per*ception, that we grasp time *per se*. And this feat of conception is a misguided one, since there is no such thing as time *per se*. (Some events happen before others; but this isn't because those events occupy certain positions in some event-independent order.)

For similar reasons, it is through conception, not perception, that we grasp space *per se*—it being events *in* space, not space itself, that we perceive—and the concept of a space devoid of events (for reasons slightly subtler than, but basically parallel to, those just stated in connection with time) is a broken one.

To sum up: the concepts of space and time are theoretical concepts. So, in trying to analyze causality in terms of the concept of spatiotemporal position, Hume is trying to understand causation in terms of two highly theoretical concepts. Thus, Hume's analysis of causality, like so many other analyses of his, is *not* empirical; for that analysis assumes that what are in fact constructs are givens. More importantly, that analysis is false; for, in taking those constructs as givens, it assumes, falsely and incoherently, that there exists an event-independent spatiotemporal order.

6.0 Hume's supposedly empirical analysis a foil for a non-empirical analysis[H13]

Hume says that we don't observe forces. But this seems false. When you are trying to lift a heavy suitcase or trying to arm wrestle a much stronger person, you are, so it would seem, observationally aware of a force.

Hume considers this very point. In response to it, he says that what you are aware of is a mere sequence of events. You have various perceptions. Some are visual—you see your opponent's sweaty, determined face. Some are tactile—you feel your opponent's increasingly sweaty hand. Some are kinesthetic—you feel your arm moving the wrong way. But Hume says you don't perceive any force in all this—even though, because you are in the grips of some wrong theories, you think you do.

But this seems very wrong. It's a datum that you *feel* your arm being forced to move in a way you don't want it to be. That doesn't mean that this feeling is accurate. Maybe it's a hallucination, like one's visual perception of a pink elephant. But it's a datum that you *have* such a feeling, even though it isn't a datum that it's accurate; and in denying that it's a datum, Hume isn't being true to his own empiricism.

Plus, this response of Hume's is decidedly contrived. If, in describing your arm-wrestling experience, you left out the part about how you *felt* as though your arm was being moved the wrong way, you'd be seriously under-describing your experience.

Thus, *if* the observational evidence is to be taken at face value, forces *do* exist. This doesn't mean that the observational evidence *should* be taken at face value. But it means that Hume's denial of causality cannot, if legitimate, be based on strictly empirical grounds. And this means that his supposedly empirical argument against causation (a) isn't empirical and (b) is only as good as some other, purely *a priori* argument.

That other argument, it will be recalled, is as follows (what follows is not a quotation):

> (NF)[339] The very concept of a force is an incoherent one. Were they to exist, forces would just be more constituents of the physical world. They would thus be no more capable than vases and rocks and so forth of constituting the cement of the universe. The idea of a force that *wasn't* that sort of cement is a

> contradiction in terms. Therefore the concept of force is an incoherent one."

We saw that vases and rocks *are* such cement. Any given physical object is a series of events any two of which are "cemented" together. Therefore, NF is no good. Therefore Hume's analysis of causality is no good.

Chapter 16 Hume’s Second Analysis of Causality

1.0 The counterfactual analysis

Hume gives two analyses of causation. We've just discussed the first. According to the second,

> (CF)[340] For E_1 to cause E_2 to occur is for it to be the case that E_2 wouldn't have occurred unless E_1 had occurred.

This is known as the ***counterfactual analysis of causality.***

Why advocate CF? The elevator comes because somebody pushed the button. If the button hadn't been pushed, the elevator wouldn't have come. The gun went off because somebody pulled the trigger. If the trigger hadn't been pulled, the gun wouldn't have gone off. No cause, no effect. Hence CF.

1.1 CF easily confused with a weaker position

CF must be distinguished from the related, but weaker claim that:

> (CF*) Given any two events E_1 and E_2, such that E_1 precedes E_2, E_1 caused E_2 to occur just in case E_2 wouldn't have occurred unless E_1 had occurred.

What is the difference between CF and CF*? A story will help us answer this question. Smith is mathematically omniscient---there is no mathematical truth that he doesn't know. But he knows nothing that *isn't* a mathematical truth. (So he doesn't know that Paris is the capital of France or that snow is white.) Thus, given any mathematical truth *T*, Smith knows *T*, and that given any *non*-mathematical truth *T**, Smith does *not* know *T**. It's therefore true that:

> (SM*) *T* is a mathematical truth iff Smith knows *T*.

But it *isn't* true that:

> (SM) *For T* to be a mathematical truth is for Smith to know *T*.

Similarly, given only that CF* is true, it doesn't follow that CF is true. But if CF is true, it does follow that CF* is true.

CF and CF*, though different statements, are ***extensionally equivalent***. In other words, given any two events E_1 and E_2, if CF entails that E_1 caused E_2, so does CF*, and *vice versa.* If CF is to be correct, it must be extensionally correct—that is, it must entail E_1 caused E_2 just in case E_2 wouldn't have occurred unless E_1 had occurred. Given that CF and CF* are extensionally equivalent, the one is extensionally correct iff the other is as well.

2.0 Why, even if it's extensionally correct, CF is false

Jones drinks poison and as a result dies at time *t*. (From now on, we'll leave the "at *t*" implicit.) It may well be true that:

(SB) Jones wouldn't have died if he hadn't consumed poison.

SB seems to entail that:

> (WB) Given any world *W* in which Jones doesn't drink the poison but is otherwise just like our world (or, in any case, is as much like our world as Jones' not consuming poison in it allows it to be), Jones doesn't die in *W*.

Question: Why doesn't Jones die in *W*? Because nothing in *W* causes him to die. No cause, no effect. In our world, the cause is present. In *W*, it isn't. WB's truth is thus to be understood *in terms of* the fact that, in our world, Jones' drinking the poison is what killed him. So SB no more constitutes an analysis of the statement "Jones died because he drank the poison" than "Smith (the mathematical prodigy mentioned in the previous section) believes that there are no regular, 17-sided polygons" constitutes a proof or an explanation of the fact that there are no regular, 17-sided polygons or of any other mathematical truth. And just as Smith's believing that there are no regular, 17-sided polygons is to be understood *in terms of* the truth of that statement, not *vice versa*, so WB is to be understood *in terms of* the truth of SB and thus *in terms of* the fact that Jones' drinking poison is what killed him, not *vice versa*. In general, "E_2 wouldn't have occurred where E_1 did not," when true, is to be understood *in terms of* the fact that E_1's occurring is what caused E_2. So even if CF is extensionally correct, it's no analysis of the concept of causality.

3.0 The vicious circularity of CF, and an outline of a positive analysis of counterfactual truth

CF assumes that the concept of counterfactual truth is not itself to be explained in terms of the concept of causality. But that may well be false. The statement:

> (SB) If Smith hadn't fallen off the building, he wouldn't have died,

seems to be equivalent, or nearly so, with:

> (SB*) Falling off the building caused an otherwise unthreatened Smith to die.

But not *all* counterfactuals register causal facts. For example:

> (SH) If six-feet tall Smith were half his current height, he'd be three-feet tall,

is equivalent, or nearly so, with:

> (SH*) Smith is six-feet tall entails that anything that is three-feet tall is half Smith's height.

Notice that SH is analytically true, whereas SB is non-analytically true. It thus seems to me that counterfactuals are just disguised causal or logical assertions and that, consequently, alternative worlds have no place in any decent analysis of them. In any case, these cases provide us with evidence that counterfactual truth is itself to be understood in terms of causality and that CF is therefore guilty of vicious circularity.

4.0 Why CF isn’t even extensionally correct

CF assumes that that there’s no causal redundancy in the world—that, were a given mechanism to fail, nothing would step into the breach. But sometimes there are back-up mechanisms. If, for some strange reason, electrocuting condemned criminal Smith didn’t kill him, the state would have found some other way to do so. Thus, the statement “the cause of Smith’s death was electrocution” isn’t equivalent with “if Smith hadn’t been electrocuted, he would not have died.” And this seems to show that CF is wrong.

Admittedly, CF has little trouble dealing with this case, and with many others of its ilk. Smith didn’t just die; he died in a very specific way. And he wouldn’t have died that way if, instead of being electrocuted, he’d been shot. This suggests that it’s only when events are seriously under-described that there appear to be counterexamples to CF.

But this maneuver doesn’t always work. *A* and *B* are both trying to push the same doorbell button. Since neither is aware of the other’s presence, each acts exactly as he would have in the other man’s absence. The button only has room for one finger. Both fingers get there at the same time, but *A*’s finger, being the bigger one, crowds out *B*’s. The requisite electrical signal is sent at time *t*. Had *A* been absent, *B*’s would have struck the same doorbell; and, given the rapidity with which *B*’s arm was moving, the electrical signal would have been sent at precisely *t*.

It may be that, were it not for *A*’s presence, *B* wouldn’t have pushed the doorbell in exactly the same *way* as *A*. (Our story may even require this.) But the signal is sent at the right time, and that’s all that matters. It doesn’t matter how exactly *B* pushed the doorbell. Doorbells are binary systems. The signal is either sent or it isn’t.

4.1 An objection

In response, it might said that:

> (LF)[341] "Perhaps the laws of physics don't validate your point," one might protest. "Maybe it's a matter of physical law that nothing happens in *quite* the same way in a world where its actual cause is absent and some stand-in has to take its place[H14] ."

If correct, LF is a complete disaster for CF. First of all, LF is an empirical point. Empirical investigation might or might not bear it out. But even if it did bear it out, the mere fact that empirical research was needed to verify LF is enough to show that LF has no place in a philosophical analysis of counterfactuals. Such an analysis is supposed to identify what, as a matter of logic, is meant by statements like SB. It's obviously an interesting question what the correct *physical* analysis of SB is. But here we're doing logical, not physical, analysis; we're making it clear what SB *means.* And surely what SB *means* doesn't include some as-of-yet undiscovered laws of physics.

We may conclude that the counterfactual analysis of causation is a failure.

Chapter 17 Counterfactuals

1.0 David Lewis on counterfactuals

Should it turn out that the concept of counterfactual truth has no logical integrity, it would follow that CF had no logical integrity. But we can't know whether that concept has any integrity until we have an answer to the question: What does it mean to say that, if *P* had been the case, *Q* would have been the case? In the next chapter we'll examine the orthodox answer to this question and, having found fault with it, replace it with one of our own.

The most famous analysis of counterfactual truth is due to David Lewis (1941–2001).[342] His analysis is beautifully clear and straightforward. According to Lewis, the meaning of[H15] :

> (KN)[343] "if kangaroos didn't have tails, they would fall over"

is:

> (OK)[344] Given any world *W* where kangaroos don't have tails, kangaroos fall over, provided that *W* is as much like our world as is permitted by the fact that, in it, kangaroos don't have tails.

In general, the right analysis of counterfactual truth, according to Lewis, is this:

> (LA)[345] "given any world *W* where *P* is the case, *Q* is the case, provided that W is as much like our world as is permitted by the fact that, in it, *P* is the case."

2.0 The role of other worlds in Lewis' analysis

Lewis' analysis obviously assumes that other "worlds" exist. This raises the question: What does Lewis mean by the term "world"? There are two options.

> Option 1: Like our world, these other worlds are space-time manifolds that are populated by objects and events.
>
> Option 2: These other "worlds" are representations of some kind or other (sets of statements, presumably).

Let's consider each option. For argument's sake, assume that these worlds are sets of statements. In that case, if LA is correct, counterfactuals merely register interpropositional relations. In other words, KN merely registers the fact that the *statement* "kangaroos don't have tails" bears some sort of relation (presumably, one of confirmation) to the *statement* "kangaroos fall over." But if *that's* the right analysis, then until this relation is identified, LA *fails* to say what counterfactuals are. And once it does identify that relation, it won't be necessary to talk about other worlds at all.

Lewis sees this and, for this very reason, he holds that the worlds in question are real worlds—space-time manifolds populated by events and objects. Lewis is thus a *realist* about other worlds.

3.0 Some consequences of modal realism

As you might predict, the standard objection to Lewis' analysis is as follows:

> (SO) We don't know if there are other worlds. What we *do* know is that, even if there are other worlds, we don't know anything about them. So nothing happening in these other worlds, if they even exist, provides the basis for our counterfactual knowledge.

[346]SO is obviously true. In any case, we'll argue for it in Section 5.0.

But let's start out by *granting* Lewis' assumption that counterfactuals, when true, hold in virtue of states of affairs that obtain in veritable alternative worlds. If we show that *given this assumption*, LA fails, then we've shown that it fails relative to its own terms and that there is thus no hope for it.

Let W be *our* world—the real world. In W, Smith intentionally shoots Jones, killing him. There were no other threats to Jones' life; therefore:

> (SK) If Smith had not shot Jones, Jones would be alive.

If LA is right, SK means the following:

> (LASK): Let *W** be a world where Smith doesn't shoot Jones but that is as much like our world as that fact lets it be: in *W**, Jones is alive.

In *W*, Smith doesn't *just* shoot Jones. That event has a cause. (It actually has a plurality of causes. But for the sake of simplicity, we'll just speak of "the cause" of that event.) Let *E* be that cause. Presumably one of *E's* components is an intention on Smith's part to kill Jones. But whatever *E* is, it can't occur in *W**, at least not in its entirety. So, in *W**, either Smith *doesn't* intend to kill Jones and therefore doesn't shoot Jones; or he does shoot Jones, but Jones is wearing a bulletproof vest. It could be anything. To fix our ideas, let's suppose that, in *W** Smith's hand seizes up just before he pulls the trigger, the gun drops, Jones gets away. Let *E** be the event that, in *W**, makes Smith's hand seize up.

What we said about *E* is true of *E**. *E** didn't just come out of nowhere; it itself had a cause. Let E^*_2 be that cause. And E^*_2 had a cause—and E^*_2's cause had a cause, and so on. So, in *W**, we have a long chain of events ($E^*_1....E^*_n$) not a single one of which occurs in *W*. Depending on how far back this chain has to go, it could make *W** *very* different from *W*. In fact, *W** might be *so* different that it might not be able to contain a Smith or Jones to begin with; and even if it did, there's no guarantee that Smith would shoot Jones or that, if he did, Jones wouldn't be wearing a bulletproof vest, etc.

Events have causal roots. Two worlds cannot differ with respect to a single event unless they differ with respect to all of the causal antecedents of those events. But, given that those go back to the beginning of time, that means that they'll differ so much that the one world becomes unable to validate the counterfactuals of the other.

It may be that *some* events are completely and entirely uncaused. But, in *W*, that happens only where micro-events are concerned. Smith's shooting Jones (in *W*) is an extremely complex event, involving innumerably causally cohesive parts and having deep roots in the past. So even if *some* events in our world are uncaused, Smith's shooting Jones isn't one of them.

And even if it *is* one of them, it's useless as far as LA is concerned. A logical analysis of LASK can't assume the results of empirical facts not in evidence and therefore cannot assume that Smith's shooting Jones is totally uncaused. In fact, such an analysis, being of a logical as opposed to an empirical kind, can't make *any* empirical assumptions.

3.1 Some consequences of modal realism (continued)

But even if it could assume this, LA would still be in trouble. First of all, we can't say that, in *W**, Smith's *not* shooting Jones is just as uncaused as the event of his doing so in *W*. In other words, we can't say that, in *W**, Smith's hand *just* seized up—that, in other words, there was no causal basis for it. For, to the extent that this event was *un*caused, any other event could have happened instead of it. We could, of course, just stipulate that, in *W**, Smith's hand seizes up. But that won't help. If, in *W**, Smith's hand *just* seizes up, then there will be at least one *other* world—call it *W*#—such that *W*# is just like *W** except that, in *W*#, Smith's hand *didn't* seize up or, if it did, something else happened that led to Jones being killed. In which case, LASK will be false. For, contrary to what LASK says, there *would* be a world satisfying the relevant conditions in which Jones dies. It's not a good idea, in general, to posit causal slack in the worlds we want to validate our counterfactuals, since that same policy makes those very worlds unreliable in the way of validating our counterfactuals.

Lewis' own way of dealing with the problem is different from those just considered. In his view, some *miracle* happens in *W** that prevents *E* from occurring there. So even though the natural laws in *W** demand that *E* happen there (or, at any rate, demand it no less than the laws of our world demand it here), a "miracle" suspends these laws, and Jones lives.

Lewis is saying that, if a world validates LASK, there is some miracle in it (at a certain time, in a certain place). So *W**, by virtue of being able to validate LASK, *also* validates the following counterfactual:

> (MLASK) If Smith had not shot Jones, then a miracle would have happened.

*W** is, by hypothesis, a world where Smith doesn't shoot Jones but is otherwise maximally like our world.

In *W**, there's a miracle (at a certain time and a certain place). So *W** makes MLASK true if it makes LASK true. But MLASK isn't a true counterfactual. And, of course, if a miracle is needed to validate MLASK, then a miracle will be needed to validate *any* counterfactual. What we just said about MLASK is true of "if Jones hadn't driven drunk, he wouldn't have crashed his car," "if Brown hadn't been so diligent in college, he wouldn't

have gotten into a reputable graduate program," and so on. So assuming these counterfactuals were true, the same would hold of (e.g.): "if Brown hadn't been so diligent, a miracle would have happened," which is obviously a preposterous counterfactual.

Also, Smith's (actually) shooting Jones is an enormously complex event, involving trillions upon trillions of mass-energy displacements. In *W**, those events can't happen. Which means that, in *W**, it isn't just one miracle that has to happen: it's trillions upon trillions of them. Which, in its turn, makes *W**'s relevance in this context extremely questionable. We're finding that, in order to get *W** to validate LASK, we have to doctor it—to mutilate it. But the very fact that we have to tamper with it—that we must employ such artifices to find a world that, in keeping with LA, can validate LASK—practically proves that LA is wrong. A theory that only doctored data supports is a bad theory, and Lewis' analysis of counterfactuals is untenable or, at the absolute least, grossly under-supported.

4.0 How Lewis' realism about possible worlds almost saves the day for the Humean

If we're right, counterfactuals are identical with causal statements. According to David Hume, nothing makes anything happen. If Hume is right, it's hard to see how anything could, in any real sense, cause anything to happen. But this seems a little strange.

David Lewis accepts Hume's analysis. This is because, like Hume, Lewis is a hardcore empiricist and, for the same reasons as Hume, rejects the idea that there are forces. But Lewis rightly thinks that denying the existence of causality is going too far. He sees that, given a realistic attitude towards other worlds along with an acceptance of the counterfactual analysis of causality, the presumption that events cause other events to occur can be reconciled with Humeanism. (By "Humeanism" I mean the view that nothing makes anything happen.) For these reasons, Lewis says that:

> (LC)[347] In our world, event *x* causes *y* to occur if, given any world *W* where *x* doesn't happen but is otherwise as much like our world as that fact permits, *y* doesn't occur.

But we've already seen why this stratagem fails. Given that *x* occurs in our world, a world in which *x* doesn't occur will bear so little resemblance to our world that it won't be able to validate the target-counterfactual (viz., "if *x* hadn't happened, *y* wouldn't have happened").

5.0 The basic problem with Lewis' analysis (LA): revisiting SO

We obviously can't see or otherwise sense-perceive what is going on in other worlds; nor can we sense-perceive evidence of what is going on in such worlds. Any knowledge that we have of what is going on in them is parasitic on our knowledge of counterfactual truths. So even if:

(i) had *x* not happened, *y* wouldn't have either,

is true, just in case:

(ii) given an *x*-free world W that is otherwise just like ours, *y* doesn't happen in W,

our *knowledge* of (i) has nothing to do with other worlds. This shows that we know (i) to be true because we know of some fact in *our* world that coincides with the fact registered by (i). This, in its turn, shows that other worlds have nothing to do with counterfactual truth. Which shows that LA is false.

Chapter 18 Explanatory vs. Statistical Probability

1.0 Humeanism and probability statements

In keeping with common sense, I believe that statements about the past confirm statements about the future and, more generally, that statements about the known confirm statements about the unknown.

It is hard to reconcile this belief with Humeanism. David Lewis is the only Humean who has even come close to succeeding in showing Humeanism to be compatible with the presumption that statements about the known can confirm statements about the unknown. To understand his argument, we must make a few preliminary points about probability.

Let's suppose that:

> (BU)[348] At time *t*, there are 10 balls in the urn. Nine are white. The other one is black. And, at *t*, there's nothing in the urn besides these 10 balls. You know all this. But this is *all* you know that's relevant to the probability of your pulling a white as opposed to black ball out of the urn.

BU isn't analytic. Nor are the statements composing it. But it *does* seem to be analytic that:

> (BB)[349] Given a single chance to pull a single ball out of the urn, the probability that you'll remove a black ball from the urn is 1 in 10.

It thus seems obvious that, *given* BU, it is analytic that BB holds.

But it isn't analytic. It isn't even true. Given *only* BU, you are no more entitled to think that you'll withdraw a white ball than you are entitled to think that you'll withdraw a black ball or a green ball or an iguana.

It may be that, given BU *along* with some reasonable assumptions (e.g., that it isn't a law of physics that white objects cannot be displaced or that, at time *t*, all white balls in urns turn into black balls or green balls or iguanas), BB is analytic.

But if any one of the parenthetical assumptions is false, you have *no* chance, given BU, of pulling a white ball out of the urn. Since these assumptions aren't analytically true, BB doesn't analytically follow from BU.

In general, when words like "probable" (and "probability," etc.) refer to

statistical probability they must be taken in a strictly non-temporal sense. The statement:

> (*) Given that, at time *t*, there are 10 apples in the barrel and only three oranges (and no other objects), your chances at later time *t** of selecting an orange are 3 in 13,

is false. What is true is:

> (*1) Given that, at time *t*, there are 10 apples in the barrel and only three oranges (and no other objects), your chances at time *t* of selecting an orange are 3 in 13, *but only* if the term “selecting” is used metaphorically.

More precisely, (*1) is true *only* if it coincides with the trivial claim that, if three out of the 13 objects in the barrel are oranges, then three out of the 13 objects in the barrel are oranges.

Of course, even a Humean can grant the truth of trivial probability statements, such as (*). So *if*, by some remote chance, it could be shown that non-trivial probability statements (e.g., given that in the past all metal has expanded when heated, it will probably continue to do so) could be identified with trivial ones, *then* the Humean *would* be able to accommodate the presumption, so crucial to thought and scientific explanation, that non-trivial probability statements are correct. David Lewis makes a noble effort to do just this.

2.0 Why for this reason Humeanism can’t explain explanation

Let’s start by describing the viewpoint of common sense. In saying that Smith will *probably* disgrace himself at his job interview, given that he drank an entire bottle of gin right before, you are saying that he all but guaranteed an unacceptable performance at the interview. In saying that Brown’s victory is probable, given that he drugged his opponent’s water, you’re saying that he all but necessitated his opponent’s defeat. In these contexts, the word “probably” doesn’t have the emptily statistical meaning borne by its homonym in statements like:

> (NC) Given that better than nine out of 10 people don’t own a pogo stick, a person chosen at random probably won’t own one.

If common sense is right, statistical probability isn’t explanatory probability (even though statistical probability is sometimes a consequence of explanatory probability: other things being equal, explanatory probabilities are expressed by statistical probabilities; and it’s from this fact that the Humean conception of probability derives what little credibility it has).

The Humean denies all this, since he thinks that nothing makes anything happen. But this means that he can't explain the fact that anything can be explained. A world where nothing makes anything happen is one where nothing has any roots—where, no matter what the conditions are, any outcome is just as likely as any other.

The Humean has three choices:

> (i) reject the presumption that some things can be explained;
>
> (ii) stop being a Humean, since Humeanism is inconsistent with this presumption;

or:

> (iii) try to reconcile Humeanism with this presumption.

Obviously Humeans choose (iii).

They defend this choice by saying that, contrary to what we said a moment ago, the concepts expressed by words like "cause" and "probable" *are* of a strictly statistical kind. "Given E_1, E_2 is probable" means "most events of the same kind as E_1 are accompanied by events of the same kind as E_2." And, as previously discussed, "E_1 causes E_2 to occur" means "given any event of the same kind as E_1, there occurs an adjacent, subsequent event the same kind as E_2."

The latter claim is a complete failure. There is no way to say what it is for events to be "of the same kind" as E_1 or E_2 that doesn't render Hume's theory of causality viciously circular. (We saw in Chapter 15.) The very same argument *mutatis mutandis* holds in connection with the Humean theory of probability. But Lewis chooses to argue for this theory; and he does so very cleverly. So let's hear him out.

3.0 Lewis' argument[350]

We've just seen that, even within a Humean framework, probability judgments are rational when they merely redescribe statistics—that, for example, (#) may be rationally accepted if it's just a way of saying that, at a given time, one person out of two weighs more than 50 pounds. Probability judgments are irrational, given a Humean framework, when they involve extrapolations. Lewis' metaphysical apparatus gives him a way of sidestepping this problem.

Bearing all of this in mind, let's suppose that, up until time *t*, whenever a rock with a mass of *m* or more, moving with a velocity of *v* or more, has collided with a window of thickness *t* or less, the window shattered upon impact. In that case, presumably, it's reasonable to judge that, the next time such a rock strikes such a window, the window will shatter. The non-Lewisian Humean is stuck; he has no way to validate this presumption. The Lewisian isn't stuck. Here's what he says[351]:

> (LA): Let C_1 be the class of worlds where, before time *t*, such windows shatter when struck by such rocks and where, after time *t*, such windows *continued* to shatter when struck by such rocks; and let C_2 be the class of worlds where, before

> time *t*, such windows shatter when struck by such rocks and where, after time *t*, such windows *didn't* shatter when struck by such rocks. C_1 has more members than C_2.

For the non-Lewisian Humean, judgments about what is likely to happen *next time* are great leaps into the unknown. For the Lewisian Humean, such judgments merely redescribe what is already there. To say that *next* time such a collision occurs the window is likely to break is to say that there are more worlds where it *does* break than where it doesn't. If it's *given* that exactly one of the cards in the deck is a King of Spades, it follows trivially that, given an arbitrary card, it has a 1 in 52 chance of being a King of Spades. Our judgments about what will happen to windows in *future* collisions are equally trivial.

3.1 Why LA fails

Any knowledge we have of what happens in other worlds after *t* will be at least as uncertain as our knowledge of what happens in our own world after *t*. There are two reasons for this. Any knowledge that we might have about anything in other worlds will be at least as uncertain as any knowledge that we have about anything in our own world. Moreover, any knowledge that we have of what happens after *t* in other worlds will be based on our knowledge, so far as we have any, of what happens after *t* in our world. So LA goes through only if it's assumed, question-beggingly, that we know what will happen in our own world's future.

3.2 Why LA fails (continued)

But there are deeper problems with LA. Here we must remember what Hume said about inductive inference:

> (HA)[352]: Given only that such and such *has* happened, it can't be assumed that such and such will *continue* to happen —*unless* it's assumed, question-beggingly, that what *has* happened will *continue* to happen. In general, given only that such and such is true of the known, it can't be assumed that it also holds of the unknown—*unless* it's assumed, question-beggingly, that the unknown resembles the known.

An analogue of what HA shows concerning inductive knowledge holds with respect to knowledge of other worlds. Given only that, in *other worlds,* such and such happens after time *t,* it can't be assumed that such and such happens after *t* in *this* world—*unless* it's assumed, question-beggingly, that what holds in circumstances already known to us also holds in circumstances not yet known to us. So LA goes through only if it's *assumed,* question-beggingly, that inductive inference works.

In objection to these points, one might say that:

> *You, the author, are missing the very point of LA. Supposing that there are 10 balls in the urn, and only one of them is white, the statement ("WB" is short for "white ball"):*
>
> > *(WB) the probability of an arbitrary one of them being white is 1 in 10,*
>
> *holds as a matter of definition and, therefore, can be affirmed without begging the question against HA. The beauty of LA is that it assimilates apparently non-innocuous probability judgments into the class of trivially true, totally innocuous probability judgments, such as WB. LA shows that:*

(UB) the probability that future unsupported objects will fall is extremely high, given that all such objects (apart from helium balloons, etc.) have fallen so far.

UB, which is non-trivial and (in this context) objectionable, is in the same category as WB, which is trivial and unobjectionable.

This line of thought embodies a confusion that has vitiated the efforts of epistemologists and philosophers of science for centuries. This confusion is of such importance that it deserves a section of its own.

4.0 Statistical vs. explanatory probability and the solution to the lottery paradox

The second word in WB is nothing more than a homonym of the second word in UB. In WB, the word "probability" has a purely statistical meaning. In UB, it has an *explanatory* meaning. Philosophers have tried to assimilate explanatory probability to statistical probability. And, while this may initially seem like good sense, it's as misguided as trying to show that a bank (financial institution) is a certain kind of bank (river's edge). This is made clear by the following paradox, commonly referred to as the "lottery paradox" ("LP," for short):

> *(LP) There's a lottery. One million tickets are sold. There's one winner. So, given any one ticket, the probability of its losing are 999,999 in 1,000,000. You buy a ticket. Let x be this ticket. You know that ("LT" is short for "losing ticket"):*
>
> *(LT) the probability of ticket x's losing is 999,999 in 1,000,000.*
>
> *But do you know that it'll lose? No. It may even have been rational for you to buy the ticket. Given what you stand to lose by buying the ticket and given what you stand to gain in the remote chance that you win, it may have been in your interest to buy it, even after you take into account that the chances of your losing is 999,999 in 1,000,000. But this decision couldn't possibly have been rational if you* knew *that your ticket would lose.*
>
> *In general, there's no knowledge where there isn't certainty. If, given the information at your disposal, there is* any *chance, no matter how small, that P isn't the case, then you don't know P. We all know (or so let us suppose) that:*
>
> *(UB) the probability that future unsupported objects will fall is extremely high, given that such objects have fallen so far.*

Is there any chance that, all of a sudden, unsupported objects will no longer fall? Yes. (The statement "all of a sudden, unsupported objects will no longer fall" isn't ruled out by the laws of logic. So it could, theoretically, be true.) Therefore, you don't know that they won't fall; and, by obvious extensions of this arguments, you don't know anything about the future (apart from sterile trivialities; e.g., "whatever happens, happens"). The very concept of "probable" knowledge is incoherent. The merely probable is ipso facto not known. To know it to be 99.9999% likely that unsupported objects will continue to fall is to not know that they'll fall.

4.1 The problem with LP

LP uses word “probability” in an equivocal manner. Once this is seen, LP loses any appearance of cogency that it might otherwise have had.

Let’s take another look at:

> (LT) the probability of ticket *x*’s losing is 999,999 in 1,000,000.

LT says that there are one million tickets but only one winner. That’s all it says. But UB doesn’t say anything of the sort. UB doesn’t make a *statistical* point. It doesn’t say: “nine times out of ten, unsupported objects fall.” It doesn’t say anything at all like that. UB makes a point about the relative merits of two competing hypotheses, these being:

> (H_1) unsupported objects will continue to fall,

And:

> (H_2) unsupported objects will not continue to fall.

UB is true if, given the data that we have, H_1’ s being true would undermine the integrity of our belief system to a lesser degree than H_2’ s being true. Given the information that we have right now:

> (SB)[353] there is only an extremely low probability that walking on stilts for an hour while wearing a beanie with a propeller on it is a foolproof cure for cancer.

Does SB say that better than nine cases of cancer out of 10 *aren’t* cured in this way? No. It doesn’t make *any* sort of statistical claim. SB is a claim, not about statistics, but about explanatory integrity. It says that more damage will be done to our belief system if it turns out that cancer *can* be cured in that way than if it turns out that it cannot. If, by some remote chance, cancer *can* be cured in that way, everything we “know” about medicine and pathology is false. But the damage wouldn’t end there. Our medical beliefs are rooted in

general assumptions about the causal structure of the world. (For example, given the information that Smith has gone into anaphylactic shock whenever he's eaten peanuts or anything with peanuts in it, even cutting-edge medical researchers will assume that Smith will go into anaphylactic shock if he eats peanuts that have been painted green. This assumption is rooted in generally held causal beliefs. No inside information is involved.) Were it found out that SB was false, many of the commonsense-based underpinnings of our medical beliefs would have to be reconsidered. But all of those beliefs remain intact if cancer cannot be cured in that way. It is improbable that walking on stilts (etc.) cures cancer to the extent that our current beliefs depend on its *not* being a cure. And it's probable that walking on stilts (etc.) isn't a cancer cure to the extent our beliefs depend on it.

In general, *P* is probable to the extent that our beliefs depend on it, and improbable to the extent that they depend on its negation.

5.0 Two kinds of probability

Counter*statistical* events are innocuous, but counter-*explanatory* events are not. Smith rolls double-sixes five times in a row. Statistically unlikely—yes. Important—no. Must we change our belief system because of it? No. Need we reconsider our assumptions about the basic structure of the world? No. And it makes no difference whether it's five double-sixes in a row or a billion. The improbable sometimes happens, and that's the end of it. When it happens, we have a nice anecdote for cocktail parties; we don't have a situation worthy of scientific scrutiny. But we *do* have such a situation if it turns out that Smith is using telekinetic powers to make the dice come up sixes. For, in *that* case, some deeply held convictions of ours have to be revised.

In describing Smith's rolling double-sixes five times in a row as "improbable," what one is saying is very different from what one would be saying in describing Smith's having telekinetic powers as "improbable." In the first case, one is affirming a mere statistical improbability. One is saying only that such things don't *often* happen. In the second case, one isn't saying anything of the sort. One is saying Smith's having such powers is inconsistent with some fundamental beliefs of ours—with beliefs that, if jettisoned, would take a lot of other obviously (or at least presumably) correct beliefs with them.

Given *any* event, there is a way of describing it that makes it improbable in the *statistical* sense. In fact, *any* given event, described at all precisely, has a statistical probability of vanishingly close to zero. I flip a coin and it lands somewhere nearby. Nothing improbable about this. But it's a safe bet that no coin has ever moved in *exactly* the way this one did after I flipped it.

Described as a case of my flipping a coin, the event isn't improbable at all. But that same event is extremely improbable if described as a case of an object's moving in manner M, where M is the exact way that it did move. But even though nothing apart from that coin has ever moved in manner M, that coin's doing so isn't a matter that deserves the slightest scientific scrutiny. Its moving in that way is consistent with everything we know about the world, and there is therefore nothing to be explained. But *there is* something to be explained if Smith turns out to have telekinetic powers, since his having such powers is improbable in the *explanatory* sense. It's irrelevant that it is also improbable in the statistical sense. *Every* event is improbable in the statistical

sense.

My winning the lottery is improbable in the statistical sense, but not in the *explanatory* sense. And that's why, before the winner is announced, we can't say that we *know* it won't win: my winning is consistent with everything we know about the world. My sprouting wings in the next five minutes is improbable in the *explanatory* sense. And that's why we *can* say that we *know* it won't happen. It's *in*consistent with everything (or much, at any rate) that we know about the world.

6.0 “Grue” and the problem of counterinduction

A puzzle devised by Nelson Goodman[354] will clarify, as well as support, what we’ve said:

> (GA)[355]: Let’s invent a new word: “grue,” to be defined as follows. Something is grue if it is green before Jan. 1, 2010, or blue thereafter. So a shirt that’s green on Oct. 12, 2007, is grue, as is a shirt that’s blue on Nov. 11, 2012. Let’s assume that, before Jan. 1. 2010, all emeralds have been green. (From now on, we’ll abbreviate “Jan. 1, 2010,” as “2010.”) In that case, all emeralds have been grue prior to 2010. Anything that is green-and-examined-before-2010 is also grue. So the fact that all emeralds examined before 2010 are green means that all emeralds examined before 2010 are grue. We are assuming for argument’s sake that, given the fact that all emeralds examined before 2010 were green, we are inductively warranted in believing that all emeralds examined after 2010 are green. By that logic, it follows that, given the fact that all emeralds examined before 2010 were grue, we are inductively warranted in believing that all emeralds examined after 2010 are grue. But anything examined after 2010 that is grue is *blue*, not *green.* So, given that all emeralds before 2010 were green, our inductive grounds for believing they’ll be *blue* after 2010 are as strong as our grounds that they’ll be green.

The degree to which a given fact is *statistically* improbable depends on how it is described. Given some emerald *x* that was examined before 2010 and found to be green, “*x* is grue after 2010” is, from a strictly *statistical* viewpoint, equiprobable with “*x* is green after 2010” and also with “*x* is blue after 2010.” But that’s no cause for alarm. Statistical improbability is innocuous. It’s explanatory improbability that you have to worry about. But the explanatory probability of “*x* is green after 2010” dwarfs that of “*x* is grue after 2010” and, therefore, “*x* is blue after 2010.” The supposition that objects don’t change color (except under conditions that emeralds don’t satisfy, at least not in virtue of making it to 2010) is deeply

embedded in scientific theory and also in the commonsense underpinning of scientific theory. If it's wrong, then science is in serious trouble, and so is human knowledge in general.

In response to this, it will be said that:

> (OA)[356] If GA is right, emeralds that "turn blue" at the beginning of 2010 *don't* change color. They stay the *same color*—grue. And emeralds that "stay green" *do* change color, since they go from being grue to having some other color. (Or, to be more precise, emeralds *do* change color at the beginning of 2010—but only *relative* to one set of color-terms ["red," "green," etc.]. But emeralds stay the same color *relative* to different color terms ["grue," "bleen," etc.].)

OA simply isn't true. Pictures of green emeralds (taken before 2010) show them *not* to have the same color as blue things after 2010. Grueness and greenness are different *properties.* But they're not different *colors*. Grueness isn't a color at all—it's a color-*related* property.

Chapter 19 The Raven Paradox[H16]

1.0 An illustration of the brokenness of Hume-based conceptions of causation and explanation: The raven paradox

Because they don't grant the existence of causality, except of a degraded and reduced kind, Humeans are categorically incapable of producing viable theories of confirmation.[357] One sign of this is that Humean theories of confirmation always have paradoxical consequences.

The most famous of these famous paradoxes is the so-called Raven paradox. This was discovered by C.G. Hempel[358], who was himself a hardcore Humean. We've already considered Hempel's unsuccessful, Humean analysis of explanation, and we'll soon consider Hempel's equally unsuccessful, Humean attempt to solve the raven paradox.

In any case, here is that paradox:

> (RP) Logically equivalent sentences are confirmationally equivalent.

This means that, if *P* and *Q* are logically equivalent, then not only does evidence *E* confirm *P* just in case it confirms *Q*, but *E* confirms *P to degree n* just in case it confirms *Q* to that same degree.

The sentence:

> (1) "All ravens are black,"

is logically equivalent with:

> (2) "All non-black things are non-ravens."

The existence of non-ravens (e.g., pianos, pants, carpet fibers) that are pink or green confirms (2), but does *not*, unless our intuitions are very wrong, confirm (1). So, impossible though it must be, (1) and (2) are logically, but not confirmationally, equivalent.

An analogous paradox is associated with each of many other pairs of logically equivalent sentences. For example:

> (i) "If it's made of metal and is being heated, then it's expanding."

is logically equivalent with:

> (ii) "If it isn't expanding, then it isn't both made of metal and being heated."

(ii) is confirmed by the fact that the non-expanding Tupperware in my freezer is not made of metal (or a fortiori of metal that is being heated). But, unless our intuitions are very wrong, that fact provides no support for (i).

Some, including Hempel himself, have argued that, our intuitions notwithstanding, the existence of pink pianos *does* confirm (1) to degree n exactly if it confirms (2) to that same degree and that the existence of non-expanding Tupperware in freezers does confirm (i) exactly as much as it confirms (ii). Others have argued that the paradox should be solved by advocating non-standard logics, relative to which logical equivalence doesn't guarantee confirmational equivalence.

We'll evaluate these solutions after we put forth our own. It's clear that, whatever merits these solutions have, they are counterintuitive. And the very idea of logical equivalence not guaranteeing confirmational equivalence is of questionable coherence. But even if it can be shown that an internally consistent logic could be invented relative to which logical and confirmational equivalence pull apart, it seems excessive to throw out classical logic to solve this one problem. Fortunately, the paradox can, I will argue, be solved without going to such lengths.

2.0 Inductions by enumeration identical with crypto-inferences to the best explanation

In this context, (i) may be treated as saying the same thing as:

> (MH) For any *x*, if *x* is made of metal and *x* is being heated, then *x* is expanding.

And (ii) may be treated as saying the same thing as:

> (HM) For any *x*, if *x* isn't expanding, then *x* isn't made of metal that is being heated.

MH can be taken *either* to express an "accidental generalization" *or* to express a principled relationship. Taken in the first way, it does nothing more than affirm a purely de facto concomitance of two properties (viz., the property of being heated metal and the property of being something that expands). Taken in the second way, MH seems to say that all heated metal expands, *the reason being* that a thing's being heated metal provides some kind of nomic basis for its expanding. In this section, we will focus on the first, strictly extensional, reading. In the next, we will focus on the second, nomically committed reading.

Let us paraphrase MH, when it is taken as a purely de facto generalization, as follows:

> (MHA)[359] For any *x*, if *x* is made of metal and is being heated, then, by happenstance, *x* is expanding.

So, in this context, take the operator "by happenstance" to be short for "not because *x*'s having the first property provides a causal or nomic basis for its having the second." We don't need to worry about whether this corresponds to the usual meaning of "as a matter of happenstance," since we are using this expression in a stipulative manner.

If we take MH in this way—if we take it to mean the same thing as MHA—does Hempel's paradox arise? No. Let phi and psi be any two properties such that *x has psi* isn't an analytic consequence of *x has phi*, or *vice versa*. If I know that there is no causal or, more generally, nomic connection between

x's being a phi, on the one hand, and *x*'s being a psi, on the other, then I have no reason to infer, from *x*'s being a phi, that *x* is also a psi. Even if, thus far, every ball withdrawn from the urn has been black, unless I have reason to believe that this concomitance has some basis in natural law or in some causal mechanism (e.g., the presence of a force field surrounding the urn that allows entry only to black balls, or some person's resolute intention to allow no non-black balls into the urn), I must *ipso facto* regard it as nothing but coincidence that, thus far, every ball withdrawn has been black; and I must therefore *not* think it more likely than not that the next ball withdrawn will be black—for to think otherwise would be to commit the Monte Carlo fallacy. [360] Thus, supposing that, in fact, this concomitance is mere coincidence, the generalization *all the balls in the urn are black* fails to receive any positive confirmation from its instances.[361] For the same reason *mutatis mutandis*, the statement *all non-black things are not balls in the urn*, unless read as having nomic force, fails to receive support from *its* instances.

Of course, it also follows that HM, if read as doing nothing more than affirming a de facto concomitance, doesn't receive support from its instances. Thus:

> (HMA) For any *x*, if *x* isn't expanding, then, by happenstance, *x* isn't made of metal that is being heated,

doesn't receive support from the fact that the non-expanding Tupperware in my freezer isn't made of metal that is being heated.

So, if MH and HM are read as having no nomic content—if they are read as doing nothing but expressing de facto concomitances—then, once it is seen that de facto generalizations don't receive support from their instances, HM doesn't even *appear* to receive support from the fact the non-expanding Tupperware in my freezer isn't made of heated metal, and MH doesn't even *appear* to receive support from the fact that the metal pan I put in the oven is expanding. A fortiori the former fact doesn't even *appear* to confirm HM more than it does MH, and the latter fact doesn't even appear to confirm MH more than it does HM.

And we therefore don't get the kind of (apparent) confirmational non-equivalence that gives rise to Hempel's paradox. For that paradox depends precisely on its being the case that, despite the (alleged) fact that HM and MH are logically equivalent, they are not, or at least appear not to be,

confirmationally equivalent. And they appear to be confirmationally non-equivalent only in so far as (e.g.) the existence of a heated, expanding metal objects confirms MH but not HM, and only in so far as (e.g.) the existence of non-expanding, non-heated, non-metal objects confirms HM but not MH. But once it is made clear that de facto generalizations don't receive support from their instances, it follows that, if MH and HM are read as free of nomic content, nothing even appears to give more support to the one than to the other. And the paradox fails to arise.

2.1 Inductions by enumeration identical with crypto-inferences to the best explanation (continued)

So if Hempel's paradox is to arise, at least one of MH and HM must be taken as doing more than affirming a de facto concomitance. And it seems that, thus taken, MH is to the effect that all heated metal expands *for the reason that* a thing's being metal that is being heated provides some kind of nomic or causal basis for its expanding. Let us paraphrase MH, when it is taken in this way, as follows:

> (MHN)[362] For any *x*, if *x* is made of metal and is being heated, then, as a matter of principle, *x* is expanding.

So, in this context, take the operator "as a matter of principle" to be short for: "because *x*'s having the first property provides a causal or nomic basis for its having the second." Echoing what we said earlier, it is irrelevant this may not correspond very well to the usual meaning of "as a matter of principle," since we are here using this expression in a stipulative manner.

At this point, we have a choice. We can take HM as having no nomic content—as taking it to have the same import as HMA. Or we can take it to have nomic content. Let's consider both options, starting with the first.

Given that MH is taken to do more than affirm a de facto concomitance, if MH is taken to affirm nothing but a de facto concomitance, the two will, quite trivially, not be equivalent. For obviously MHN is not equivalent with HMA. The first is committed to its being the case that *inter alia* a thing's being made of heated metal provides some kind of nomic basis for—or, in any case, is nomically or causally connected in some way with—that thing's expanding. The second isn't committed to there being any kind of nomic, or otherwise not purely extensional, connection between the property of not expanding and the property of not being made of heated metal. MHN is true only if certain nomic connections hold; HMA can be true in the absence of those connections. So there is no equivalence, or even the appearance of equivalence for that matter.

We've just seen that, if we take MH to have nomic force (which we must —see above—if Hempel's paradox is to arise), Hempel's paradox doesn't arise if we *don't* take HM to have nomic content. In other words, if we take

HM to do nothing more than affirm a de facto concomitance, then it won't, as we've just seen, even appear to be equivalent with MHN, and nothing even resembling Hempel's paradox arises. So if that paradox is to arise, then we must take HM to have nomic content.

In light of this, consider the sentence:

> (HMN) For any *x*, if *x* isn't expanding, then, as a matter of principle, *x* isn't made of metal that is being heated.

HMN says that a thing's not being an expander provides some kind of (partial) nomic basis for its not being heated metal. Of course, the relationship that HMN bears to HM is analogous to the relationship that MHN bears to MH. But what is most important here is that HMN would be the right way—or, in any case, the (or, at least, *a*) natural way—to take HM *if* the latter were taken to do more than affirm a de facto concomitance between two properties.

To clarify this last point, suppose that you are speaking to a friend who is about to drink a large quantity of arsenic but who doesn't realize that consuming arsenic in large quantities leads to death. You tell him, warningly, "people who consume large quantities of arsenic die soon thereafter." Here you are obviously not affirming a purely extensional connection between, on the one hand, the property of being a person who drinks large quantities of arsenic at a given time and, on the other hand, the property of being something that dies soon thereafter. What you are saying is that a person *x*'s drinking large quantities of arsenic at a given time *leads to*, or in any case provides some kind of nomic foundation for, *x*'s dying soon thereafter, *that* being why there is a concomitance between the property of being a person who drinks arsenic, on the one hand, and the property of being dead soon thereafter, on the other. That this is the right interpretation is made clear by the fact that you would be *failing* to warn your friend not to drink the arsenic if you had instead said: "All people who have drunk arsenic died soon thereafter. But that had nothing to do with their drinking arsenic. You see, it turns out that the only people who've ever drunk arsenic belong to religious sect *X*; moreover, all of those people were about to be executed. One of the precepts of *X* is that, if you are about to be executed, you must drink arsenic in order to get into heaven. And that's why, thus far, each person who's consumed arsenic died soon thereafter."

The most natural way to take HM to have nomic force is to take it to mean that a thing's being a non-expander provides some kind of causal or nomic basis for its not being heated metal. And that is precisely the reading of HM that is embodied in HMN.

HMN is clearly *not* equivalent with MHN. Suppose, if only for argument's sake, that a thing's being made of heated metal *causes* it to expand. So suppose that:

> (MHC) "as a matter of natural law, *x*'s being heated metal *causes x* to expand,"

which would certainly be sufficient for the truth of MHN. In light of this, consider the statement that:

> (HMC) "as a matter of natural law, *x*'s not expanding causes *x* to not be metal that is being heated."

Suppose it is a law that *x's being a human who drinks arsenic at t causes x to be dead at some later time t**. This very reasonable and (give or take a few nuances) correct statement obviously doesn't entail, and therefore isn't equivalent with, the false and incoherent statement that *x's not being dead t* causes x to be a human who doesn't drink arsenic at earlier time t.*

For similar reasons, MHC is not equivalent with HMC. So, since MHC and HMC are sufficient for the truth of, respectively, MHN and HMN, it follows that MHN and HMN are not equivalent: they don't hold in the same models, the same possible worlds. In fact, whereas MHN, or something similar to it, *does* hold in our world, the same doesn't go for HMN or anything at all like it.

2.2 Inductions by enumeration identical with crypto-inferences to the best explanation (continued)

We've seen, first, that MH must be read as having some kind of nomic content if Hempel's paradox is to arise (i.e., if there is to even be the appearance that MH is logically, but not confirmationally, equivalent with HM). And we've also asserted that, if MH read as having nomic force, its meaning is given by MHN, and we made the same assumption *mutatis mutandis* about HM.

But, it must be admitted, we have provided no argumentative basis for this assertion, our grounds for it being purely intuitive.

There is another possible nomic reading of MH:

> (MHW)[363] "Thanks to the existence of some kind of nomic (or causal) connection between the two properties about to be mentioned (viz., the property of being heated metal, on the one hand, and the property of expanding, on the other), for all *x*, if *x* is heated metal, then *x* is expanding."

How is MHN different from MHW? MHN says pretty much the same thing as:

> (*) For all *x*, if *x* is heated metal, then *x* is expanding, and this is because *x*'s being heated metal provides a nomic basis for *x*'s expanding.

By contrast, MHW says that:

> (^) For all *x*, if *x* is heated metal, then *x* is expanding, and this is because *either x*'s being heated metal provides a nomic basis for *x*'s expanding *or x*'s expanding provides a nomic basis for *x*'s being heated metal.

To be sure, the expression "provides a nomic basis for" is an obscure one, which I've done little or nothing to clarify; and it might therefore seem that there is no difference between (*) and (^). But, obscure though that expression may be, (*) and (^) are demonstrably different, and non-

equivalent, statements.

It's true (or so let us suppose, for argument's sake) that all people who consume large quantities of arsenic die soon thereafter. And this is not an accident: human biology being what it is and the chemical properties of arsenic being what they are, there's a nomic connection between consuming arsenic and dying. Vague as the expression "provides a nomic basis for" is, would it be correct to say that being a dead human at 3:10 P.M. "provides a nomic basis for" being a person who drinks arsenic at 3:00 P.M.? More generally, supposing that *x*'s having psi at *t** is an *effect* of *x*'s having phi at earlier time *t*, does it make sense to say that *x*'s having psi at *t** provides a nomic basis for *x*'s having phi at *t*? No. My being obese provides a nomic basis for my having a heart-attack. But surely my having a heart attack doesn't provide a nomic basis for my being obese. And that's

all we need to know in this context about what it means to say that one thing provides a "nomic basis" for another. So let us proceed.

Suppose that, in possible world *W*, all things made of heated metal expand. Further suppose that, in *W*, *x*'s being something that is expanding causes *x* to turn into something that is metal whose temperature is continuously rising and whose volume is constantly increasing. And, to make things simple, suppose that, in *W*, all things made of heated metal got that way by, at some point in their histories, being things that were expanding. In that case, MHW would be correct in *W*: for there would, in that world, be a nomic connection between a thing's being heated metal, on the one hand, and its being something that expands, on the other; moreover, the fact that, in *W*, all heated metal expands would have a basis in this nomic connection. But it would not, at least not on that account, be correct to say that, in *W*, a thing's being made of heated metal provides a nomic basis for its expanding. In *W*, things expand *before* they become expanding heated metal. So, in *W*, a thing's being heated metal obviously cannot *cause* it to expand.

But there is a further point: in *W*, a thing's expanding is not to be *understood in terms of its being heated metal*. It is the other way around. In *W*, a thing's expanding is causally *and explanatorily* more basic than it's being heated metal. In any case, just as (see above) it would be wrong, or at least very odd, to say that being dead at *t** provided a *nomic basis* for drinking arsenic at earlier time *t*, it would, for parallel reasons, be wrong to say that, in *W*, being heated metal provided a nomic basis for being a thing that expands. And this entails that, however vague the expression "provides a

nomic basis for" is, * and ^ are, very clearly, *not* equivalent.

3.0 This argument concluded

First of all, MHW obviously isn't logically equivalent with HMA: the former does, whereas the latter does not, allege the existence of certain nomic dependencies. More importantly, MHW isn't equivalent with:

> (HMW) "thanks to the existence of some kind of nomic (or causal) connection between the two properties about to be mentioned (namely, the property of not expanding, on the one hand, and the property of not being heated metal, on the other), for all *x*, if *x* is not expanding, then *x* is not heated metal."

Given only that there is a nomic connection between a thing's having phi and it's having psi, it doesn't follow that there is a nomic connection between a thing's *not* having psi and it's *not* having phi. There is a causal, and therefore a nomic, connection between Smith's falling off of tall building *X* at *t*, on the one hand, and his being dead at a later time *t**, on the other. Smith's falling off of *X* is an event. It is something that makes things happen (e.g., various bones of Smith's being shattered); and, what is probably a more general notion, it is something in terms of which things can be explained.

Compare Smith's situation with Brown's. Brown does *not* fall off *X*, or any other building, at *t*; and Brown is alive and well at *t**. Obviously Brown's *not* falling off *X* at *t* doesn't *cause* Brown to be alive at *t**. More generally, Brown's being alive at *t** is not, except perhaps in some extremely derivative sense, to be explained in terms of Brown's not falling off of *X*. Whereas Smith's falling off of *X* is an event, Brown's *not* falling off of *X* is *not* an event, and it doesn't consist of events. A chair consists of events (various miniscule displacements of mass-energy), and so does a corporation (various actions on the part of sentient beings); and so, indeed, does any enduring entity—any structure, any "object," any non-instantaneous event. But Brown's *not* falling off *X* isn't identical with an event and isn't constituted by events.

Of course, since it's a truth, and therefore a "fact," that Brown did not fall off of *X*, we might say that the (negative) "fact" of his *not* falling off of *X* exists no less than the (positive) fact of Smith's falling off of *X*. But that doesn't undermine our point that Brown's *not* falling off of *X* isn't identical with, or constituted by, events.

While it is a matter of intense debate what nomic relations are, it is pretty clear they hold only among occupants of the space-time manifold. They clearly hold among events (my pounding the table causes it to break). Maybe they hold among things other than events. Maybe, for example, they hold among conditions. And maybe they hold among objects. But while it is unclear whether *all* spatiotemporal entities stand in nomic relations, it *is* clear that nothing *non*-spatiotemporal stands in them. My pounding the table is nomically connected to the table's breaking. But my *not* destroying Jupiter isn't nomically related to anything.

My pounding the table is an event; it makes things happen; things can be explained in terms of it; it has a place in the tapestry that constitutes the spatiotemporal world. My *not* destroying Jupiter is nowhere in that tapestry, and therefore isn't fit to be nomically related to anything. There is no event of my not destroying Jupiter. There is thus no condition or object that is constituted to any degree by that non-event. Therefore, it isn't identical with, or constitutive of, anything that falls within the scope of natural law. It isn't, therefore, nomically connected to anything.

A thing's being composed of metal that is of a certain temperature is, in the relevant respects, like my pounding the table. Things are to be explained in terms of it. And, though not itself an "event," it is *constituted* by various interacting events, all of which occupy space-time. So the skillet's having a certain temperature, at a given time, is something that can, and indeed obviously must, stand in nomic relations. And the same is true of the skillet's expanding during a given interval.

But a thing's *not* expanding is, in the relevant respects, like my *not* destroying Jupiter, and so is a thing's *not* being heated metal. So, whereas a thing's being heated metal is, or at least could be, nomically related to its expanding, it doesn't follow, and it isn't the case, that a thing's *not* expanding could be nomically related to it's *not* being heated metal.

These points make it clear why MHW isn't logically equivalent with HMW. A thing's being heated metal *can* (and, in fact, must) be nomically related to other things, and the same is true of a thing's expanding. But a thing's *not* expanding, like my *not* destroying Jupiter, isn't nomically related to anything, and neither is a thing's *not* being heated metal. Which is why, even though MH and HM are equivalent when they're read as merely affirming *de facto* concomitances they cannot always be intersubstituted *salva veritate.* In particular, they cannot be so intersubstituted in contexts

governed by the operator that begins each of MHW and HMW (viz., "'thanks to the existence of some kind of nomic (or causal) connection between the two properties about to be mentioned"). There is nothing mysterious about this. Many operators don't permit intersubstitutions of logically equivalent expressions. Logically equivalent sentences can be truth-preservingly intersubstituted in contexts governed by (e.g.) "it is true that," but not, at least not categorically, in contexts governed by "Timmy believes that." The operator "there is a nomic connection between the facts expressed by the antecedent and consequent of the following statement" doesn't tolerate intersubstitutions of logically equivalent sentences. For, even if:

> (T_1) "There is a nomic connection between the facts expressed by the antecedent and consequent of the following statement: if *x* is heated metal, then *x* is expanding,"

is true, it doesn't follow that:

> (T_2) "There is a nomic connection between the facts expressed by the antecedent and consequent of the following statement: if *x* isn't expanding, then *x* isn't heated metal,"

is also true. In fact, T_2 is false, for the reasons just given.

4.0 Hume-internal solutions to the raven paradox: Quine's solution

Quine's (1977) solution is to say that logical equivalence doesn't guarantee confirmational equivalence. This idea is totally incoherent. *P* and *Q* are equivalent if they are true in exactly the same worlds. Thus, supposing that *P* and *Q* are equivalent, it immediately follows that, given any proposition *R*, the likelihood of *R*'s being true in a world where *P* is true is identical with the likelihood of *R*'s being true in a world where *Q* is true. Quine's proposed solution is thus a non-starter.

4.1 Hume-internal solutions to the raven paradox: Hempel's solution

Hempel proposes a more plausible solution. It obviously *seems* as though the existence of a pink piano provides less support for:

(1) "All ravens are black,"

than does the existence of a black raven. And it obviously *seems* as though the existence of pink pianos provides more support for:

(2) "All non-black things are non-ravens,"

than it does for (1). But, says Hempel, these intuitions are misleading. Under ordinary circumstances, if we know that something is a piano, we almost certainly *already* know that it isn't a raven.

But imagine the following. You have bet your friend Smith $1,000,000 that all ravens are black. There is some object *x* such that, although you and Smith don't yet know it, *x* is a piano. All you know about *x* is that it's pink. In that case, when you *discover* that *x* is a piano, and thus a non-raven, you (but not Smith) will breathe a sigh of relief; and your confidence in (1) will surge, and so, of course, will your confidence in (2).

Now imagine a different scenario. You have bet your friend Smith $1,000,000 that all non-black things are non-ravens. There is some object *y* such that, although you and Smith don't yet know it, *y* is black. All you know about *y* is that it's a raven. When you discover that *y* is black, you will breathe a sigh of relief; and your confidence in (2) will surge, and so, of course, will your confidence in (1).

Thus, to Hempel, the confirmational disparities between (1) and (2) are illusory. When we imagine ourselves trying to confirm (2), we assume that we are looking at things that we *already* know to be non-ravens, but we don't make the same assumption *mutatis mutandis* when we imagine ourselves to be trying to confirm (1).

Hempel's solution is better than Quine's, but it's still no good. The problem is that, if it's known that there is no nomic or causal relationship between a thing's being a raven on the one hand, and its being black on the other, then there is *ipso facto* no reason to believe that the next raven you see will be black. Maybe you've seen a million ravens, all of them black. It's

doesn’t matter. Unless you have reason to believe that a thing’s being a raven is connected somehow to its being black, you’re guilty of the gambler’s fallacy if, on the basis of that data, you believe that the next raven you see is (likely to be) black. But if you *do* believe that a thing’s being a raven is connected in some way to its being black, then your confidence in (1) is derivative of your confidence in:

> (1*) A thing’s being a raven provides a basis of some kind for its being black.

We’ve already seen why. And we’ve already seen why (1*) isn’t equivalent with:

> (2*) A thing’s being non-black provides some kind of a basis for its being a non-raven.

If the existence of a black raven is to provide any legitimate support for (1), there must be a causal or nomic connection between ravenhood and blackness. Once this is realized, two things immediately become obvious. First, Hempel’s solution is spurious, since it presupposes that black ravens can confirm (1) *without* there being any connection of the sort just described. Second, (1) must be read as having nomic force. Since, thus read, it is no longer equivalent with (2), the paradox vanishes.

5.0 Summary of this chapter's main points

If read as doing nothing more than affirming de facto concomitances:

(MH) if *x* is made of heated metal, *x* expands,

And:

(HM) if *x* isn't expanding, *x* isn't made of heated metal,

don't receive support from their instances—and don't even appear to receive support and are therefore vacuously confirmationally identical. So they are logically *and* confirmationally identical when read in this way, and Hempel's paradox doesn't arise. If they are read as doing *more* than affirming de facto concomitances, they must be read as alleging, or at least being committed to there being, some kind of nomic connection between the properties in question. But, no matter what nomic readings we give these sentences, they don't even appear logically equivalent. (This is, to put it very crudely, because logically equivalent sentences cannot, at least not in general, be intersubstituted *salva veritate* in contexts governed by nomic operators; e.g., "there is a nomic connection between the following statements.") Bottom line: once it is realized that accidental generalizations don't receive support from their instances, this being the prerogative of principled generalizations, it isn't possible for Hempel's paradox even to arise.

Part V Explanation

Chapter 20 Scientific Explanation

1.0 Hume's regularity analysis of causation: The awkward constraints it places on would-be explainers of explanation

According to Hume, for *x* to cause *y* is for the sequence consisting of *x* and *y* to be an instance of a regularity (provided that *x* immediately precedes *y* and is also adjacent with it). We found there to be not so much as a sliver of truth in this view. But philosophers to this day accept it. Presumably, this is only because no significant alternative to it has been put forth. Given how wrong Hume's analysis seems to be ("you're pulling the trigger had nothing to do with the gun's firing"), it's hard to believe that anyone would accept that analysis unless they felt they had no other choice.

In any case, supposing it correct, Hume's analysis of causality has very definite consequences as to what it is to *explain* events. As a result, there is no shortage of Humean theories of explanation; in fact, such theories still predominate. Echoing what we said a moment ago, this is presumably because no viable alternatives to those theories have been proposed, given how wrong those theories seem to be (as we're about to discover) and given how wrong they are easily proven to be (as we're also about to discover).

1.1 Two preliminaries: The nature of scientific truth and the Duhem-Quine thesis

Statements are either true or they aren't. There are no "scientific" statements, and there are no "unscientific" statements.

To the extent that an utterance lacks a clear meaning, it constitutes a failed attempt to make a statement and therefore *isn't* a statement. And to the extent that an utterance has a clear meaning, it's either true or false, and therefore cannot meaningfully be described as "scientific" or "unscientific." If *T* is true, science has to accept it; if *T* is false, science has to reject it. Period.

That said, there are distinctly scientific ways of *generating* statements. Given some statement *S*, assuming that:

(SS) ‹*S* is a scientific statement›,

doesn't say the very same thing as:

(OS)[364] ‹*S* is a statement›.

SS must say *S* was generated in a certain way—that it was generated in accordance with procedures whose truth-conduciveness has been definitively established, or some such.

Nonetheless, a number of philosophers have argued that there is some property that all and only "scientific" statements have. For example, in his (1931) book *The Logic of Scientific Discovery*, Karl Popper (1902–1994) argued that a statement is scientific iff it is ***falsifiable.*** To say that a statement *S* is falsifiable is to say that it could in principle be shown to be false on strictly observational grounds. (The sort of falsificationism that Popper advocated is not to be confused with the very different doctrine, discussed in Chapter 1, according to which a statement is *meaningful* iff it is falsifiable.)

To be sure, *some* theories can be falsified on strictly observational grounds. Suppose that, according to theory *T*, anyone who takes medication *M* will be acne-free within five hours. Given that acne-ridden Smith took *M* at noon, it follows that *T* is false if Smith still has acne at 5:00 P.M. Given that Smith *does* still have acne at 5:00 P.M., and given that this fact is painfully obvious to anyone who looks at Smith, it follows that *T* can be falsified on strictly observational grounds.

But *T* is not representative of most theories in this respect. Whether or not a given theory has a given observable consequence is likely to depend on whether each of various other theories is correct. (To confirm a given astronomical theory, one must assume the truth of various theories concerning the behavior of light. In making this assumption, one is assuming that the instruments used to authenticate those theories were reliable ones. In making *this* assumption, one is, for reasons similar to those just given, assuming the truth of physical theories additional to those already mentioned. And so on and so forth.) So given only that some observable event hasn't occurred, it's not likely to follow that this or that specific theory must be false.

1.1.1 The Duhem-Quine thesis

Aware of the fact that theories can't typically be falsified on strictly observational grounds, W.V.O. Quine (1908–2001) took the position that *any* theory is consistent with *any* body of observational data. Given any theory and any set of observations, Quine says, there is way of modeling those observations that is consistent with that theory. Quine refers to this thesis of his as the "Duhem-Quine thesis."[365]

The so-called Duhem-Quine thesis is an extremely facile one and is easily refuted. As before, suppose that, according to theory *T*, anyone who takes medication *M* will be acne-free within five hours. Given that Smith's acne is going strong six hours after taking M, there's nothing we can say that isn't brazenly *ad hoc* and that doesn't create vast numbers of otherwise non-existent anomalies. Should an opponent of M point out this fact, his attempt to refute M would be a success, as it would have met the relevant standards of proof. For, when evaluating a scientific theory, the relevant standard of proof isn't given by the question "given the sensory data at our disposal, is it *logically* impossible for that theory to be correct?", and is instead "given the sensory data at our disposal, would countenancing that theory require us to renounce fruitful theories which there are no independent reasons to reject and which, therefore, we couldn't reject without repopulating the universe with anomalies that we had long since solved?"

It's obvious and trivial that, given any theory *T* that concerns the external world, no matter what observations we've had; it's *logically* possible that *T* is false. For, as Quine points out, it's theoretically possibly that everyone is always hallucinating and, therefore, that the data on which any given scientific theory is based is spurious. But, since this platitude is all Quine gives us in the way of substantiating his thesis, he's effectively done nothing to substantiate it.

In any case, unless it's completely hollow, the Duhem-Quine thesis can't just say that, for any theory *T* that concerns the external world, no body of observations analytically entails that *T* is false. It must therefore be to the effect that, given any body of observations and given any theory *T*, the explanatory costs of accepting *T* are no greater and no smaller than the explanatory costs of rejecting *T*.

But *that* statement is very obviously false. The data being what it is, the explanatory costs of holding onto *T* far outweigh the explanatory gains of

doing so. If we hold onto it, we're stuck with various anomalies that we wouldn't otherwise be stuck with. If we reject it, we're not stuck with those anomalies; nor are we stuck with any other anomalies that we weren't already stuck with. So depending on how it's interpreted, the Duhem-Quine thesis is either trivial or false.

2.0 Scientific explanation: Hempel's theory

Let us now turn our attention to Carl Hempel's important theory as to the nature of scientific explanation. Hempel's theory is important because it is the clearest, purest expression of the Humean conviction that all probability is statistical probability and also because of the consequent fact that, given the enormously powerful hold of Humean ideas on methodological thought, all theories of explanation can be understood in terms of Hempel's theory—either as versions of it or as reactions *to* it. Hempel's theory is known as the "deductive nomological" account of explanation:

> (DN) To explain an event *E* is to show that its occurrence is a logical consequence of the statement formed by conjoining statements describing known natural laws with statements describing the conditions under which *E* occurred.[366]

I drop an otherwise unsupported object. The laws of physics being what they are, the object must fall, given the circumstances. And to explain that event is to identify known natural laws that, when conjoined with a correct description of the circumstances, entail that the object will fall.

2.1 Hempel's analysis evaluated

What is meant by the term "law"? One possible answer is: A law is what is described by *any* true universal generalization—that is, by any statement of the form "all phi's are psi's." Few have explicitly given this answer. But we'll now see that Hempel has no choice but to accept that answer, ridiculous though it is.

If the word "law" is taken in this way, DN has the consequence that the following counts as an "explanation" of the fact that the pencil on my desk is yellow[367]:

> (NE) Premise #1 (description of the relevant natural law): Everything is such that it's either yellow or it isn't a pencil on JMK's desk.
>
> Premise #2 (description of initial conditions): *x* is a pencil on JMK's desk.
>
> Conclusion (thing being explained): *x* is yellow.

But NE isn't an explanation of the fact that my pencil is yellow. The problem, of course, is that[368]:

> (YP) Everything is such that it's either yellow or it isn't a pencil on JMK's desk,

doesn't express a natural law. But why not?

Because, says Hempel, it *doesn't support counterfactuals.* The statement:

> (ME) Metal expands when heated,

expresses a law of nature because a piece of metal that isn't being heated *would* expand *if* it were heated. Nothing comparable holds of YP. Given an arbitrary pencil, YP doesn't entail that it *would* be yellow *if* it were on my desk. In general, says Hempel, laws are expressed by universal generalizations that support counterfactuals.

Hempel is surely right that a statement must support counterfactuals if it is

to express a natural law. But DN collapses if laws are understood in this way. Let *m* be some unheated metal object. Why is it that *m would* expand *if* it were heated? Because, the laws of nature being what they are, heating *m* would *cause* it to expand. So the fact that *m would* expand if heated is to be understood *in terms of* the fact that ME expresses a natural law.

Remember what we said earlier about counterfactuals: they are thinly veiled causal statements. "Smith would be alive if he hadn't jumped" means: "jumping is what caused the death of an otherwise unthreatened Smith." So far as that causal claim is warranted, it is because there is known to be a law that, given the circumstances, demanded Smith's death. The counterfactual claim is to be understood in terms of a causal claim, and the causal claim is to be understood in terms of a statement describing a natural law. DN thus becomes viciously circular if "law" is defined as "what is expressed by a counterfactual-supporting universal generalization."

Causal claims are *themselves* explanations. In being told that lung cancer is what led to Smith's death, I'm being given an explanation, albeit an incomplete one, as to why he died. Counterfactuals, as we know, are identical with causal claims. It is only where explanations are *already* available that counterfactual claims are warranted.

By the same token, if we don't know that it was arsenic consumption that killed Smith—or, if not Smith, then somebody else—we have no basis for affirming the counterfactual that Jones, who isn't consuming arsenic, *would* die *were* he to do so.

Thus, we are warranted in affirming counterfactuals only where we *already* have explanations. So to understand the concept of law in terms of the concept of counterfactual truth *is* to understand the law in terms of the concept of explanation. Thus, if laws are understood in this way, DN is analyzing the concept of explanation in terms of itself.

So long as the concept of law is understood in terms of the concept of explanation, DN is viciously circular. Is there some other way to understand the concept of law? This question exposes a serious problem with DN. Hume tried to show that laws are *mere* concomitances. He denied the existence of causal ties; so he couldn't take laws to be descriptions of such ties. And he therefore couldn't take "metal expands when heated" to describe the *connection* that we presume to hold between *x*'s being heated metal and *x*'s expanding. He had only one option: the law isn't some structure or metaphysical glue *underlying* the fact that the property of being heated metal

is always found in conjunction with the property of expanding; the law *is* that regularity. Laws *are* the regularities that non-Humeans typically regard as their *expressions.*

The advocate of DN is forced to take this position. So long as he thinks of laws as being given by statements that describe *structures* of some kind—the structures that are manifested by regularities—the advocate of DN is analyzing the explanation in terms of itself. For it isn't *until* there are explanations, that there can be any good reason to posit such structures. The structure is posited to validate preexisting explanations. This makes it viciously circular to understand laws in terms of causality. This, in turn, makes it viciously circular to understand laws in terms of counterfactual truth, given that counterfactual truths are to be understood in terms of causality. If his position is to be free of such circularity, the advocate of DN must take laws to be mere concomitances. In his view, the law that all metal expands when heated *is* the regularity that, in the eyes of a non-Humean, merely expresses that law. In general, for the advocate of DN, laws are regularities—nothing more.

But in that case, *any* true universal generalization expresses a law. YP ("everything is such that it's either yellow or it isn't a pencil on JMK's desk") describes a regularity. It correctly states that nothing has the property of being a pencil on JMK's desk on March 11, 2009, unless it also has the property of being yellow. But it doesn't express a law.

If laws are identified with regularities, as opposed to the structures that (so non-Humeans think) underlie them, then DN ends up having the same problems as Hume's analysis of causation. People often know what caused what, even though they couldn't for the life of them identify the relevant generalizations. It often happens that, *in order* to identify the relevant generalization, one must know what caused what, and therefore that one must *already* have an explanation of the event in question, if only a crude one, if one is to identify the relevant generalization. Jones turns the ignition in his car, and his car explodes. Smith witnesses this and wants an explanation for what he just saw. He thinks about the sequence of events that he just witnessed. Jones got into his car; he put the key in the ignition and then turned it; the engine turned over a few times; at that point, the car exploded. If Smith falsely believes that well-known evil person Brown used his telekinetic powers to blow up the car, Smith's search for covering laws will be futile. It is only if Smith *does* have a correct explanation (*Jones' starting*

the engine set off a bomb that, presumably, somebody deliberately installed in the car), if only an extremely low-resolution one, that he can even begin to search in a meaningful way for the relevant law.

Laws are hyper-explanations.[369] They are explanations of the explanations. First there are the *anomic* explanations. ("Nomic" means "involving or having to do with natural law." "Anomic" means "not involving natural law.") The explanations that we come up with in everyday life are anomic—"heating it caused it to expand," "falling caused him to break his leg," "drinking arsenic caused him to die," "being made of pinewood caused it to float when placed in water." The laws explain these explanations. "There is an underlying structure that guarantees that a thing will expand if it's made of metal that is being heated/that a person will break his leg as a result of such and such a fall/that a person dies if he consumes arsenic/that a thing floats if it's pinewood in water."

So Hume and Hempel have it backwards; for they believe that the event is to be explained in terms of the law. To be sure, once the law is in place, new and better explanations are possible. But those higher-level explanations stand on the shoulders of lower-level ones that model the data without the help of statements describing laws. Hypotheses concerning laws are distillations of existing, low-level explanations.

In conclusion, DN is viciously circular and is therefore a failure.

2.2 More on the circularity inherent in the Hempel-Hume analysis of explanation

To explain an event is not to show that it is an instance of a general regularity. First of all, given *any* event *E*, it's possible to contrive regularity of which *E* is an instance. Right now, my pencil, which is yellow, is occupying region R on the surface of my desk. No other object has ever occupied that exact region; and, we will suppose, no other ever will occupy it. So every object that ever has or will occupy *R* is yellow.

Consider the following three pairs statements:

1. *x* is yellow.
2. *x* occupies *R*; *every occupant of R is yellow; i.e., it is a regularity, to which there are no exceptions, that every occupant of R is yellow*.

a. *x* just expanded.
b. *x* is made of metal and *all metal objects expand when heated*.

i. *x* is floating.
ii. *x* is made of pinewood and pinewood floats.

Bearing these three statement-pairs in mind, consider the following three questions:

1*: Why is *x* yellow?
a*: Why did *x* expand?
i*: Why is *x* floating?

2 is clearly *not* an answer to 1*. But why? Because the regularity indicated by the italicized part of 2 is obviously of the wrong kind. But why is it of the wrong kind? Because a thing's being in region R obviously doesn't *cause it* to be yellow.[370]

Obviously 2 isn't the sort of explanation that advocates of DN have in mind. They have in mind explanations of the form:

2#: *x* has phi and *every instance of phi is yellow*,

where the italicized part indicates a *relevant* regularity. But, as we've just seen, the underlined part indicates a *relevant* sort of regularity *only* if a thing's being phi *causes it* to be yellow. But if we know that a thing's being a phi *causes* it to be a psi, then we *ipso facto* already have a wealth of explanations, if only crude ones, at our disposal. For if we know that *O* is causally responsible for *O#*, then we *ipso facto* have an explanation of *O#*.

So *relevant* regularities---regularities of the sort that DN-advocates obviously have in mind ("DN" being short for "deductive nomological analysis of explanation")---*presuppose* the existence of explanations. Thus, DN is either viciously circular or, at best, it provides a correct analysis only of explanations of a derivative and therefore relatively unimportant kind.

The exact significance of these remarks will become clear, as will the justification for them, when we consider a* and i*, which we will do right now. The italicized part of 2 obviously *does* correspond to the sort of regularity that advocates of DN have in mind---but *only* because it's *already* known that a thing's being made of heated metal *causes* it to expand. Somebody who doesn't *already know* that *x*'s being made of heated metal caused it to expand won't be able to identify the relevant covering law. But somebody who *does* know that *x*'s being made of heated metal caused it to expand will have no use for that law. Such a person *already knows* that *x*'s being made of heated metal caused it to expand; so such a person *already* knows that a thing's being made of heated metal causes it to expand. So unless a person *already* knows why *x* expanded, he won't be able to find the right covering law. But if he knows why *x* expanded, he already knows the relevant covering law. So a person can't identify the sort of explanation that advocates of DN have in mind *unless* they *already* have an explanation of the fact for which an explanation is needed and *unless*, therefore, they have no use at all for the DN-explanation of that event.

3.0 Psychological explanation: the glaring failure of Hempel-Hume conception of explanation to provide a reasonable account of it

Some explanations concern psychological events: they concern thoughts, feelings, intentions, etc. along with the overt behaviors in which they result. When we consider such explanations, it becomes clear that what Hume and Hempel have to say about explanation fails, in a very ugly and obvious way, to give a correct account of psychological explanation; and, in the process of seeing why this is so, it will become even more clear than before why exactly that analysis fails to give a correct account of *any* sort of explanation.

Consider the following two statements, both of which we'll assume to be correct:

> Statement#1: Smith goes to church every Sunday.
> Statement#2: All people as devout as Smith go to church every Sunday.

Here is what advocates of DN say[371]:

> (VDN) To explain the fact that Smith goes to church every Sunday is to identify some pair of characteristics, phi and psi, such that it is already known every instance of phi is an instance of psi and such that Smith is already known to be a phi. Smith is already known to be devout, and it's already known that people as devout as Smith go to church every Sunday. Therefore, Statement #2 (henceforth "S_2") provides an explanation for the fact described by Statement #1 ("S_1").

VDN is obviously not the right explanation of Smith's behavior. The right explanation is this: he's devout; his being devout, along with some beliefs of his as to how a devout person ought to act, prompt him to go to church every Sunday. Supposing that everyone as devout as Smith goes to church every Sunday, that regularity is a *consequence* of facts about the causal structure of the world. Far from being something *in terms of which* Smiths behavior is to be explained, that regularity is *itself* to be understood in terms of the behavior

of people like Smith. The fact that devout people regularly attend Church is parasitic on the fact that Smith's being devout causes him to go to church every Sunday, along with the fact that Brown's being devout causes him to go church every Sunday, along with the fact that Johnson's being devout causes her to go to church every Sunday, etc. So, in effect, one explains S_1 in terms of itself, and thus explains nothing, if one explains S_1 in terms of S_2. For exactly similar reasons, one explains heated copper object *x*'s expansion in terms of itself, and thus doesn't explain it at all, if one explains it in terms of the fact that copper objects expand when heated.

4.0 The beginnings of an explanation of explanation

One principle always underlies the search for answers (here I am referring exclusively to attempts to answer *empirical* questions—questions concerning the spatiotemporal):

(MC)[372] causal anomalies must be minimized.

For everything to be explained is for every cause to be known.

"But surely that isn't true," it will be said. "I can know *what* killed Tim (he ate fish) without knowing *how* it did so. (Was the fish rotten? Was it poisoned? Was Tim allergic to fish?)" Yes, but knowledge of the how is knowledge of the what. How-knowledge is high-resolution what-knowledge. If all you know about the cause of Tim's death was that it was brought on by his eating fish, then you don't *really* know what caused his death. Your not knowing the "how" consists in your not knowing the specific mechanisms involved; and your not knowing those mechanisms consists in your not knowing the "what."

That having been settled, let us turn to the question: How are causal anomalies minimized? The answer:

(ED)[373] The elimination of causal anomalies consists of the elimination of discontinuities.

You see a red ball roll in one end of a tunnel with opaque walls. A moment later, you see a similar ball roll out the other end. Given this information, it is natural to suppose that:

(H_1) The ball you saw entering the tunnel rolled from one end of it to the other and then exited it.

It's a theoretical possibility that H_1 isn't the right explanation. Maybe what you saw is part of a magic-trick. But other things being equal, H_1 is almost certainly correct. And even if it isn't correct, it's obviously a very reasonable hypothesis. The reason is that it eliminates the discontinuities in the events that you observed by supplementing them with events that you did

not observe. Your observations tell you that some ball was instantaneously annihilated and that some other ball instantaneously popped into existence a few feet away. But if H_1 is correct, none of these *apparent* discontinuities is an actual discontinuity. At least part of what makes H_1 a good hypothesis is that, to the extent that it's correct, reality is continuous. This suggests that to explain is to eliminate discontinuities. Explanation is discontinuity-elimination.

In response to this, it will be said that:

> (OH)[374] We can use your own example to prove that, contrary to what you are alleging, explanation is not the same thing as discontinuity-elimination. A ball rolls into one end of a tunnel and a similar ball rolls out the other. But the second ball isn't identical with the first. The first was obliterated the moment it was completely inside the tunnel. At that instant, an iguana suddenly materialized in the place previously occupied by the ball. That iguana ran from one end of the tunnel. But just before it would have emerged from the other end of the tunnel, the iguana hit a force field. Because of that collision, the iguana turned into a ball just like the one that entered the tunnel and, in addition, that ball rolled out of the tunnel.
>
> Let H_2 be this hypothesis. H_2 eliminates each of the discontinuities eliminated by H_1. But whereas H_1 is a good hypothesis, H_2 is rubbish. So you're wrong: to explain is *not* to eliminate discontinuities.

OH is misguided. Supposing that H_2 is correct, a red ball suddenly vanishes and is replaced with an iguana that, although continuous with the ball in respect of its *position*, is discontinuous with it in respect of its morphological, chromatic, and chemical properties; and, a moment later, that iguana suddenly vanishes and is replaced with a ball that, although continuous with the iguana in respect of its position, is discontinuous with it in respect of its shape, color, and chemical properties. Thus, merely H_2 replaces observed discontinuities with unobserved discontinuities. H_1 doesn't posit any of the discontinuities posited by H_2. So contrary to what OH says

there are many discontinuities that H_1 eliminates that H_2 fails to eliminate. Moreover, there are discontinuities that H_2 creates that H_1 does not create. These facts are not only consistent with, but strongly support, our thesis that explanation is discontinuity-elimination.

That said, there is a gap in our defense of that thesis:

> (OH*) Clearly erroneous hypotheses can do just as good a job as correct ones of eliminating discontinuities. Here is an example. You see a ball roll into an opaque tunnel. Immediately upon entering the tunnel, the ball quickly but *continuously* turns into an iguana. The iguana runs to the other end of the tunnel. Once at the other end, the iguana quickly but *continuously* turns back into a ball; and that ball rolls out the other end of the tunnel.
>
> Let H_3 be this hypothesis. Even though it's obviously unacceptable, H_3 does as good a job of eliminating discontinuities as H_1. Thus, contrary to what you are saying, discontinuity-elimination cannot be the essence of explanation.

Contrary to what OH* alleges, H_3 does *not* do as good a job as H_1 of eliminating discontinuities. First of all, even though it seems, at least initially, to do as good a job as H_1 of eliminating spatiotemporal discontinuities, H_3 creates a number of *theoretical* discontinuities that H_1 does not create. There is no independent evidence that balls turn into iguanas or *vice versa.* So, in supposing that H_3 is correct, we're supposing *either* (i) that the physical laws governing this one situation are very different from those governing other situations *or* (ii) that the observational data that we've had thus far, which strongly suggests that iguanas *don't* turn into balls or *vice versa,* is very misleading as to the nature of reality. If (i) is the case, the laws of nature, and therefore the very structure of reality, had to undergo an abrupt and unprecedented change in order for the changes described by H_3 to be possible. (And given the reasonable, though debatable (not to mention

vague), supposition that the laws of nature are veritable spatiotemporal structures, it follows that H_3 fails to eliminate discontinuities in a particularly egregious way.) If (ii) is the case, there is so little conformity between the data of observation and the structure of reality that we have no good reason to expect any of our observation-based beliefs, including our belief in (ii) itself, to be correct. Given obvious extensions of these points, it's clear that H_2 creates similar theoretical discontinuities.

But H_1 doesn't have any of these problems. It doesn't create any of the theoretical (or spatiotemporal) discontinuities that either H_2 or H_3 creates, and it eliminates each of the discontinuities eliminated by either. All of which suggests that, just as we proposed, discontinuity-elimination is the essence of explanation.

5.0 Probability as an explanatory, not a statistical, concept (revisited)

At time *t*, a prominent medical researcher says:

> (S_1) "it's probable, but not certain, that Socrates had four kidneys."

Given all of the data available up until time *t*, S_1 is the rational thing to say.

More data becomes available. The researcher considers it. And, at time *t**, that same researcher says:

> (S_2) "Socrates only had two kidneys."

Given all of the data available up until time *t**, S_2 is the rational thing to say.

Question: Given that S_2 is true, does it follow that S_1 was false? No. Given the data available to the researcher at *t*, it *was* probable that Socrates had four kidneys. Like all probability statements, S_1 describes a relationship (one of bearing) that holds between some body of data and some hypothesis. S_1 does *not* describe some state of affairs. If it did, it would be to the effect that Socrates had two kidneys *to a certain degree*, in which case it would be meaningless.

5.1 Probability a property of statements, not of events or states of affairs

Probability is a property of statements, not of states of affairs.[375] It's a property of statements to the effect that a certain hypothesis is given a certain amount of support by a certain body of data. So it's a property of statements that describe bearing-relations. True probability statements correctly describe such relations; false ones incorrectly describe them.

An illustration: Smith tells his friend Brown that he's going on a tour of Europe. But he's vague about where he's going to go, and when. But he's not *completely* vague; he gives Brown some relevant information. A month after Smith leaves, Brown tells a friend of his that:

(PS_1) It's probable that Smith is in France right now.

PS_1 is correct. It's correct because, given the data available to Brown, it's far more reasonable to believe that Brown is in France than it is to believe that he isn't. And PS_1 is correct even if, at the time, Brown is Spain. It doesn't matter whether Brown is in France or not. What matters is that, given the available data, it's more reasonable to believe that he is than it is to believe that he isn't.

PS_1 says that, *given* such and such data, it's probable that Smith is in France. It's irrelevant that, from a phonetic viewpoint, there's no expression of the form "given such and such data." From a logical standpoint, there is such an expression (unless PS_1 is completely meaningless).

In general, probability is inherently relational. For this reason, the words "*P* is probably the case" may be correct when they come out of one person's mouth and false when they come out of some other person's mouth. This will happen if the data available to the one person makes *P* more belief-worthy than not-*P* and the data available to the other makes not-*P* more belief-worthy than *P*.

A complete description of the universe would contain no probability statements. Given any statement *S* (that isn't a probability statement), such a description would contain either *S* or not-*S*. Thus, probability is but "a measure of ignorance," as Laplace put it. Somebody who knew everything could never correctly describe a proposition as probable but not certain.

5.2 Probability not subjective

Probability is a mind-independent logical relationship holding among statements. The degree to which a given body of data *D* confirms a given statement *S* has nothing to do with anyone's opinion about anything: *D* confirms *S* to that degree whether or not anyone knows it—whether or not anyone exists. Probability statements are *useful* only to the ignorant. (An omniscient being wouldn't have to bother with them.) But that doesn't mean it's only insofar as people are ignorant that they're true. It means the opposite. They're useful to the ignorant because, when correct, they describe relations that actually hold between what (little) they know and what they'd like to know.

Given a certain body of data, it's more reasonable to believe *P* than it is to believe not-*P*, even though, given that same data, it is *not* reasonable to believe *P*. To believe *P* is to be certain that *P* is correct. The data one has may make *P more* worthy of belief than not-*P* without making *P* worthy of belief. Given the data available to me, it's more reasonable to believe that my friend Sally is in Washington than it is to believe that she isn't. But since, given that data, it's only marginally more likely than not that she's in Washington, it would be irrational of me to *believe* that she he is.

5.3 Is there such a thing as non-epistemic probability?

According to some, probability is sometimes a property of events, as opposed to statements. Here's the idea. At time *t*, electron *e* jumps to a higher orbit. Though the laws of nature didn't require it to jump, or not to jump, at that time, they did *dispose* it to jump. Here, it is said, we are dealing with a case of intrinsic, as opposed to evidential, probability. Supposing that one knew everything there was to know *other* than the fact that *e* jumped orbit at *t*, one could not on that basis predict with any certainty that *e* would jump at that time. Under these circumstances it might be said, it *would* be correct to say, that for reasons having nothing to do with ignorance on anyone's part, the statement "it is more probable than not that *e* will jump at *t*" is correct.

Suppose for argument's sake that there are indeed undetermined events—events that don't have to happen but do---one could argue that intrinsic probability is a kind of evidential probability:

> (IP)[376] Let *S* be a true statement alleging the occurrence of event *E*. *S* has an "intrinsic" probability of greater than 50% iff S_1. . . . S_n make it more likely than not that *S*, where S_1 . . . S_n is a list of every true statement *that doesn't* logically entail *S*. The idea is that, if you knew everything there was to know *other than the fact that E jumps orbit at t*, you could not learn more that was relevant to whether *E* would jump without actually learning whether *E* would jump. But, it will be said, the story is very different with statements that as a matter of natural law either must happen or cannot happen. Let *S** be a true statement falling into this category. No matter how small the likelihood that *S** is false, there are true statements that, if conjoined with the truths already known, will shrink that likelihood even further *without eliminating it.* (It may be that, because of interference effects, it's impossible to acquire more information that is relevant to whether *S** is correct or not without actually learning whether *S** is correct or not. But that would only be because those interference effects made it impossible to learn the truths that would enable you to shrink your margin of doubt.) Thus, intrinsic probability is a kind of evidential probability. One is dealing with intrinsic probability

> when, setting aside problems relating to interference effects, there *is* a limit to how much, short of learning whether the target statement is true or not, one's margin of error can be reduced, and one has non-intrinsic evidential probability if there is no such limit.

Nonetheless, I'm not convinced that intrinsic probability is, in fact, a form of evidential probability. As before, let S be a true statement alleging the occurrence of some event *E* that has an intrinsic probability of greater than 50% but less than 100%, and let $S_1 \ldots S_n$ be a list that contains every true statement other than those that entail *S*. The statement "given $S_1 \ldots S_n$, *S* is more probable than not" does not say that, given $S_1 \ldots S_n$, we're stuck with more questions than answers if we deny *S*. Were it supposed true that *Hamlet* was written by a levitating cat that, despite being incapable of self-consciousness, was deeply neurotic, all manner of beliefs concerning physics, biology, and psychology would have to be revised. But, assuming the truth of $S_1 \ldots S_n$, it doesn't much matter whether *S* is accepted or rejected. It is, by hypothesis, a free wheel. I must therefore conclude that intrinsic probability, if it exists, is not evidential probability.

I leave it to the reader as an exercise to say why intrinsic probability is not, contrary to what some might hold, identical with statistical probability; and I also leave it to the reader to say what exactly intrinsic probability is.

5.4 Explanatory vs. statistical probability (revisited)

In this section, we'll put forth some more reasons to think that explanatory (evidential) probability is not statistical probability. It will help if we start with a summary of our findings thus far.

Other things being equal, a hypothesis is to be preferred to its negation if and only if more causal anomalousness is created by rejecting it than by accepting it. The degree to which a hypothesis is to be preferred to its negation is directly proportional to the degree to which it does a better job than the other of minimizing the net amount of causal anomalousness in the world.

Probability is not a statistical notion. The word "probability" has a statistical meaning. This is the meaning that it has in the statement: "the chances of a person who's picked at random having red hair are one in eight." All this means is that one person in eight has red hair.

But this isn't the sort of probability that is relevant to explanations. When you say ("SNS" is short for "Smith no show"):

> (SNS) "Given that Smith hasn't shown up yet, he's probably not going to show up at all,"

you are *not* making a statistical statement; you are not saying that in more than 50% of situations like the present one Smith doesn't show up. SNS could be true even if Smith has never failed to show up to anything in his life. And it could be false even if, in the past, he's failed to show up to 90% of his appointments.

Statements affirming the probability of hypotheses seldom even *appear* capable of statistical interpretations. You can know that *x* is likely to do *y* in circumstances *z* without ever having seen *x*, or anything much like it, do *y*, or anything much like, in circumstances *z*, or anything much like them. For example, we have a pretty good idea of what at least some of the consequences of detonating a hydrogen bomb over the South Pole would be, even though nothing much like that has ever occurred.

"But that's where you're wrong," contemporary Humeans say. "You see, it's only if it's interpreted in an absurdly narrow way that that the statistical conception of explanation has any of these problems." And they defend this position by saying that ("SE" stands for "statistical explanation"):

(SE) If a hydrogen bomb were detonated over the South Pole, it would by no means be unprecedented. True, we haven't detonated hydrogen bombs *over the South Pole.* In fact, we haven't detonated *hydrogen* bombs anywhere. But we've detonated weapons *similar* to hydrogen bombs over places *similar* to the South Pole. Given this information, we have a pretty good idea of what the consequences would be of detonating a hydrogen bomb over the South Pole.

And while you're right that SNS may be correct even if Smith has never missed an appointment in his life, you're wrong to conclude that explanation is not to be understood in statistical terms. If one is to be justified in accepting SNS, one doesn't need to know that *Smith* ever failed to show up to anything; and one doesn't need to know that what the relevant individuals, whoever they might be, failed to do was to show up. One need only know that people similar (in some respect relevant in the present context) to Smith failed to do something similar (same qualification) to showing up. One need only know that people *like Smith* (e.g., people who have serious drug problems) failed in some way that would be comparable to Smith's failing to show up (e.g., those people did not keep their promises).

It's easy to ridicule the statistical conception of explanation by showing that events *exactly like the one being explained* haven't occurred. But the situation being explained needn't be a perfect replication of any past situation. It need only be *like* past situations in some very abstract respect.

The problem with this response is that there is no way to understand the concept of "relevant similarity" in statistical terms; and in deciding what is relevantly similar to what, the advocate of SE is drawing on hypotheses that can't possibly have received any confirmation of a purely statistical kind.

Any two things are similar in some respects and dissimilar in others. Some of these similarities are explanatorily significant, and others are not. It often happens that, despite our not having first conducted context-specific research, we can distinguish the relevant from the irrelevant forms of similarity. For

example, it's obvious that, when trying to figure out what the consequences would be of detonating a given bomb over the South Pole, it doesn't matter whether that bomb is green as opposed to red, but it *does* matter whether it's a nuclear as opposed to a non-nuclear bomb. So it's obvious that, in the relevant respects, a red hydrogen bomb would be much more similar to a green hydrogen bomb than it would be to a red non-nuclear bomb.

But this wouldn't be obvious at all if our evidence for it were strictly statistical. Given that, for any two things, one can contrive some respect in which they are similar, it follows that, with a few trivial qualifications, anything can confirm anything to any degree.

This is what Nelson Goodman's argument shows. (See Section 7.3). Find *a* respect in which green emeralds are similar to blue emeralds, and the fact that, thus far, all emeralds have been green confirms that, after this point, they'll all be blue. (All green emeralds examined before the present are similar to blue emeralds examined in the future in that they're all "grue.")

Another example: Find *a* respect in which herbivorous brontosauruses resemble cars that aren't affected in any adverse way by high-speed collisions, and the fact that all brontosauruses were herbivorous confirms that a speed collision won't damage your car. (Herbivorous brontosauruses are similar to cases of your crashing your new, as of yet unscathed, car in that they are both "carbivorous." Something is "carbivorous" if it's either an herbivorous brontosaurus or a car driven by you after the present time that's involved in a high-speed collision and isn't damaged by it.")

We know that terms like "grue" and "carbivore" pick out explanatorily hollow properties, and that's why we don't base our driving habits on our beliefs about the dietary habits of dinosaurs. We know what to infer and what not to infer *only* insofar as we know what kinds of similarity count. A *precondition* for making the right inferences is knowing what the relevant kinds of similarity are.

Thus, there's no way that judgments as to what the relevant forms of similarity are could have a strictly statistical basis. (That said, they obviously can and do have a *partly* statistical basis. But that's not enough to make a go of the statistical model.) The concept of probability is to be understood in terms of the concept of explanatory integrity, not that of statistical frequency. The concept of explanatory integrity is to be understood in terms of the concept of *causal* integrity. The concept of causal integrity is to be understood in terms of the concept of continuity of change. One very

important kind of change is change of position (motion). But not all change is change of position. Changes in temperature aren't changes in position. Nor are changes in color, shape, or structure. That said, at least some of these forms of change are to be understood *in terms of* the concept of change of position. But the same can't be said of other explanatorily important kinds of change. It can't be said of higher-order changes (changes in the manner of change), since higher-order changes aren't *spatiotemporal* changes at all. (A second-order change is a change in a spatiotemporal change; it is not itself a spatiotemporal change, and the same is obviously true of all other higher-order changes.) Earlier we considered some clearly erroneous hypotheses that seemed at first to do just as good a job of preserving continuities as some clearly worthy hypotheses. Had these initial impressions been correct, our analysis of explanation would have been in trouble. But they weren't correct. When we took a second look, we found that the clearly wrong hypotheses did a much worse job than their meritorious counterparts of dealing with higher-order continuities. And the continuities in question were ones that we already knew to be explanatorily important; they weren't contrived ad hoc to buttress our theory.

6.0 A crucial fact about scientific explanation: Reality-description methodologically subordinate to data-modeling.

There is more to explaining an event than showing it to be continuous with events in its vicinity. Let M_1 and M_n be two mental events. M_1, we'll suppose, is Smith's suddenly realizing that he's not a good composer and M_n, we will suppose, is Smith's realizing that his compositional efforts over the last 20 years have been in vain. There are, as we know, good reasons to believe that M_1 and M_n are identical with brain events---let B_1 and B_n be those events. But we'll see that, in order to explain M_n in terms of M_1, it is neither necessary nor sufficient that we show, for each i ($1 \leq i \leq n$), that B_i is continuous with the brain-events in its spatiotemporal vicinity. Scientific disciplines (e.g., psychology, physics) are individuated, as we'll see, not (merely) by *what* they study, but by the data in terms of which the objects of their study are given to them. Those data are expressed by propositions.

The same is true of events: they too are known to us through propositions. [377] Propositions consist of concepts. Disciplines model data. It isn't the event *per se* that is explained; it is some proposition describing its occurrence. Any given event---in other words, any (re)distribution of mass-energy---can be correctly described in many ways. There are many true propositions describing it.

A consequence of these points is that, for any given (re)distribution of mass-energy, there are many possible objects of explanation. Given some event *E*, let *T* be some proposition describing *E*'s occurrence that we wish to explain. Remember that it is *T*, not *E per se*, that we are trying to explain. It would be meaningless to say "I'm explaining event *E*---but there is no truth that I am explaining." It is (alleged) truths that are explained. Facts and events are explained only in the indirect sense that true propositions about them are explained.

Moving on: Unless the continuities that we posit in the hopes of explaining *T* are expressed by propositions consisting of concepts belonging to the same family of concepts to which those constituting *T* belong we haven't established continuity within our *models* of events. And since it is only through these models that we know what the transphenomenal facts are, and thus what the transphenomenal continuities are, we haven't, under the circumstances described, succeeded in establishing any sort of continuity that

is of any use in discovering the facts that lie on the other side of the data.

So the essence of explanation lies not *merely* in the discovery of continuities, but in the discovery of propositions that, because of the structures of the concepts out of which they are constructed, establish what we might refer to as “model-internal” (or “data-internal”) continuities. We will clarify, illustrate, and demonstrate the truth of these points in the upcoming chapter. Moreover, we will do this in connection with our discussion of why it is that recent attempts to “scientize” the discipline of psychology are broken and *ab initio* doomed to failure.

Chapter 21 Psychological Explanation and Interdisciplinary Reduction

1.0 Some introductory remarks

"The brain is an organ just like another." This is a popular saying among contemporary psychologists, especially among those who claim to be doing psychology in a "scientific" manner. So far as this slogan means anything, it means that psychology should be studied physiologically, not psychologically.

In addition to being tautologically false, this position embodies a failure to grasp certain basic truths about knowledge and scientific methodology. It will be the purpose of this chapter to state those truths and to draw attention to some of their more obvious consequences. One of those consequences is that the mind must be understood on its own terms. So long as psychologists treat psychological truths as low-resolution physiological truths, and so long as they don't treat the class of psychological truths as *sui generis*, psychology will be a broken and incoherent discipline.

Before we can substantiate these claims, we must make some general points concerning the nature of scientific inquiry.

2.0 What it is to be a datum

In this context, any references to intentions are to intentions that don't embody any incoherent beliefs.

Even though, when framing a hypothesis, one's *ultimate* objective is to identify the realities with which that hypothesis is concerned, that cannot possibly be one's immediate objective. For it could be one's *immediate* objective to describe those realities only if one knew what they were and therefore didn't need the help of hypotheses to acquire this knowledge. Thus, one's immediate objective must be to model the data, not to describe data-transcendent realities.

A related point is that it is *truths* that are explained. Not objects. Not events. Rocks aren't explained. Neither are migraines. "But surely explain events," it will be said. "A few years ago, my husband had a seizure. We went to a physician, looking for an explanation, and we got one." If the *truth* that your husband had a seizure had the same spatiotemporal boundaries as the seizure itself, it would have ceased to exist the moment the seizure ended, and it wouldn't have existed outside the place where the seizure occurred. Thus, the physician you later saw would have been explaining a non-existent truth. In other words, there would have been no truth that he was explaining. But there was such a truth, and there still is. That said, is no place that it is occupying. It isn't propping up a tent or heating up water otherwise behaving like things that occupy places. There isn't some place such that, by virtue of being in it or right next to it, one is better able to see or touch or hear or otherwise be affected by that truth. So it makes no sense to say that it has a location. So that proposition, just like its negation, is a non-spatiotemporal entity. But that event did have a location. So what the physician explained was a truth, not an event, and the same is true of everything that can possibly be explained. We'll see that this seemingly technical point must figure prominently in any decent explanation of explanation.

2.0.1 What it is to be a datum (continued)

Data are not raw facts. Data are propositions. Propositions are not raw facts.

When trying to say what distinguishes a given discipline from every other discipline, the relevant question is *not* "which things and which events does that discipline study?" but rather "which propositions does that discipline assume to be true and which propositions is it trying to model?"

Of course, there are some very generic propositions that all disciplines take for granted (e.g., if a thing is, in its entirety, in a given place at a given time, it cannot, in its entirety, simultaneously be in some other place that is a million miles away"). But given any discipline, there will be some class of propositions that it alone assumes to be true; and the subject matter of that discipline is identified when this class's membership is identified.

There is a plausible objection to this position:

> (PO)[378] But astronomy studies stars; geology studies landmasses; zoology studies animals. Surely, contrary to what you say, disciplines *are* individuated by what they study. Disciplines are therefore *not* individuated by which propositions they assume to be true. They are individuated, I repeat, by which objects and events they study.

Much of this is correct. For example, the facts studied by astronomy clearly don't even overlap with the facts studied by zoology.

But we mustn't let these truths lead us astray. Astronomical facts are microphysical facts. Any movement of a star is nothing but a giant constellation of microphysical facts. Two universes that are indistinguishable at the microphysical level are necessarily indistinguishable at the astronomical level. Two such universes, indeed, are indistinguishable *tout court.* But astronomy is not a department of microphysics. In fact, those disciplines are disjoint, or nearly so.

Similarly, even though all zoological facts are microphysical facts, zoology is disjoint from microphysics, or nearly so. Also, all of the facts studied by zoology are microphysical facts, and so are all of the facts studied by astronomy.

To be sure, there is no *specific* fact that falls within the scope of both

disciplines. But that's irrelevant. At the microphysical level, a dog is not so different from a comet. The general principles instantiated by the one class of microphysical facts overlaps heavily, if it doesn't coincide with, those instantiated by the other class. But that does nothing in the way of causing zoology to merge with astronomy. The fact that the class of things studied by astronomy is disjoint from the class of things studied by zoology is, we must conclude, a consequence of the fact that there are no raw facts described by data belong to both disciplines or by theoretical extensions thereof.

Thus, disciplines are individuated primarily, and sometimes solely, by the bodies of data that they are trying to model. Disciplines are individuated only secondarily, and sometimes not at all, by the facts that they describe.

PO embodies a failure to take into account that it is truths, not raw facts, that are explained. The raw facts in virtue of which astronomical facts hold are of the same ilk as those in virtue of which zoology holds. So, were it raw facts (and raw events), that were the foci of scientific explanation, astronomy would be zoology. But astronomy isn't zoology.

"But," it will be said, "when explaining star-constitutive events, one must sometimes *apply* microphysical principles in ways in which one needn't apply them when explaining animal-constitutive events." True, but irrelevant. The way in which one must apply architectural principles to build a 33-story building doesn't precisely coincide with the way in which one must apply those same principles to build a 29-story building. But it would be absurd to say on that account that, if *X* designs a 29-strory building, he is not an "architect" in the same sense as someone who builds a 33-story building.

"But," it will also be said, "the classes of *specific* raw facts studied by those disciplines are disjoint." True, but irrelevant. Given two medical researchers *X* and *Y*, it's unlikely that they study precisely the very same tissue samples. But it would be absurd to say on that account that *X* and *Y* are practitioners of different disciplines. It could very well be that *X* and *Y* are studying the very same problem, in the very same way.

2.1 What it is to be a datum: conclusion

Given only that somebody studies what is *in fact* a plant, it doesn’t follow that he’s a botanist. If his interest in those plants extends only as far as their ability to furnish him with information as to the structure of a certain kind of particle, then they are just heaps of particles to him, and it is as a physicist, not a botanist, that he is interested in them. Similarly, given only that somebody studies heaps of molecules, it doesn't follow that he’s a microphysicist; for if his interest in those heaps extends only as far as their ability to furnish him with distinctly botanical knowledge, he is a botanist, not a microphysicist.

By parity of reasoning, given only that somebody studies what are *in fact* neurological events, it doesn't follow that he's a neurologist. And given only that somebody studies what are *in fact* mental events, it doesn't follow that he's a psychologist. What determines whether somebody is a psychologist as opposed to a neurologist is a function of which truths he's trying to explain, not which raw facts those truths describe.

3.0 Psychological explanation vs. non-psychological explanation

We thus have the following facts to work with:

> (i) Science tries to eliminate anomalies. (Anomalies are not irregularities. Anomalies are discontinuities. It is in terms of concept of continuity that the concept of an explanatorily significant regularity is to be understood. See Chapter 17.)
>
> (ii) Science doesn't *directly* deal with the realities it aspires to describe. (If it did, it wouldn't have to *aspire* to describe them.) Science deals directly with data.
>
> (iii) What distinguishes discipline *A* from discipline *B* is that the class of propositions whose truth *A* assumes doesn't coincide with the class of propositions whose truth *B* assumes.

Bearing these three facts in mind, consider the following situation. Barry is a botanist. Barry leaves plant *X* in a place where, for hours on end, it is exposed to blistering sunlight. Unbeknownst to Barry, *X* belongs to a species whose members can live only in cool, shady places. After a few hours, *X* dies.

It's clear to Barry's supervisor that the cause of *X*'s death was over-exposure to sunlight. But the supervisor knows Barry to be a brilliant scientist and he is more interested in learning from Barry than in reprimanding him. The supervisor thus asks Barry to produce a maximally thorough *botanical* explanation of *X*'s death.

Assume (*per impossibile*) that Barry has complete knowledge of the microphysical activity constitutive of *X*'s being placed in the sun, remaining in the sun, and then eventually dying, and also that everything Barry says is correct.

Barry gives a thorough and precise description of each and every last micro-event involved in the series of particle-flurries mediating between, and including, *X*'s being placed in the sun and *X*'s dying. He uses no botanical or biological terms. The only non-generic terms that he uses belong to microphysics and mathematics.

Did Barry explain *X*'s death? No. We must remember that it is *truths* that are explained, not (re)distributions of mass-energy. *Things* aren't explained. Rocks cannot be explained. Propositions *about* rocks are explained. *Events* aren't explained. Truths that hold in virtue of events are explained.

Barry did an excellent job of describing the (re)distributions of mass-energy mediating between *X*'s being put in the sun and *X*'s dying. True---the first particle-flurry happened to be identical with a certain, still-living plant; and the last particle flurry happened to be identical with a ***genidentical***, recently-deceased plant.[379] So Barry successfully explained a microphysical truth that happened to hold in virtue of the same *raw fact* as the botanical truth that he was asked to explain. But he did not explain that botanical truth itself. So he failed to explain what he was asked to explain.

3.1 An objection

Some will feel that, by virtue of giving a maximally precise and thorough account of the microphysical activity involved in *X*'s death, Barry has given a maximally good explanation of *X*'s death. Such thinking is off the mark.

The *botanical* changes that mediated between *X*'s once healthy state and its eventual demise can be realized by infinitely many different *micro*physical changes. So Barry's excursion into microphysics may have explained why *this* plant died. But it did nothing in the way of explaining the deaths of plants whose situations were botanically indistinguishable from *X*'s. Had Barry given a cogent *botanical* explanation of *X*'s death, he would *ipso facto* have explained why events *botanically* similar to *X*'s being moved lead to events *botanically* similar to *X*'s dying. And by failing to give a cogent botanical explanation of *X*'s death, Barry failed to say anything about why plants botanically similar to *X*, when placed in botanically similar situations, end up dying. Therefore, Barry failed to provide a *botanical* explanation of the death of even this one plant.

3.2 Why concomitances are not enough for explanation

If there turned out to be a way to characterize botanical concepts in microphysical terms, *then* Barry's answer might constitute a *bona fide*botanical explanation of *X*'s death.

But there is no such way; nor could there be. The relevant distributions of micro-particles couldn't be described *without using botanical terms.* A given type of botanical event can be realized by infinitely many different events, no two of which have anything in common at the microphysical level. Let *K* be the least comprehensive class containing every instance (past, present, future, and possible) of some pine-tree's dying.

Obviously *K*'s membership is infinitely large. But that isn't important. What is important is that, from a microphysical perspective, *K*'s membership is infinitely diverse. In other words, there is no one microphysical property (or finitely large set of microphysical properties) had by everything in *K* and by nothing not in *K*.

But let's set aside talk of the merely possible; let's focus on the class of all *actual* pine-tree deaths. Let K* be *that* class. The likelihood that there is some microphysical property had in common by all and only K*'s members is small (but not zero). Obviously pine trees have similar cellular structures. But cells don't fall within the scope of microphysics. One reason for this is that, although cells are small in absolute terms, they are enormous relative to electrons, quarks, etc. Another far more important reason is that the concept of a cell no more helps the (micro)physicist discover principles relevant to his field than does the concept of a theocratic dictatorship.) But even if this slim possibility materialized, Barry *still* would have failed to explain X's death. Some science fiction will make it clear why this is so.

Imagine a world in which round objects were green and square ones were red and in which we knew that colors and shapes were thus coordinated. In such a world, our knowledge of the truth that *x is a green wheel* could be very useful, as could our knowledge of the truth that *y is a red wheel*: we'd know that, other things being equal, *x* would be a much better wheel than *y*. And, supposing that we lived in a world in which the microphysical truths "tracked" botanical truths in a way similar to the way in which, in the world just described, chromatic truths "tracked" geometrical truths, our knowledge of that fact could potentially be very useful.

Knowledge of concomitances. But it mustn't be thought that knowledge of

concomitances is anything more than a surrogate for genuine understanding. The next section will develop what we've already said in defense of this claim.

4.0 The concept of psycho-physical reduction

In this context, when talking about the “physical condition" of some object *X*, I mean what would be given by a description of the constellation of micro-particles composing *X* that was put forth by an infallible physicist who was acting in his capacity as a physicist.

A plant’s botanical condition ***supervenes*** on its physical condition. In other words, given two plants that are physically indistinguishable, it is impossible, as a matter of logic, that they should be botanically distinguishable. Since any given plant is composed of nothing but micro-particles, everything that happens to a plant coincides with a change in the inter-relations, if not the individual compositions, of those micro-particles. So where there are no micro-physical changes, there are no botanical changes; and where there are identical micro-physical changes, there are identical botanical changes.

It has been assumed that, because the botanical supervenes on the microphysical, it must be possible to "reduce" botany to microphysics---that is, to show that what botany has to say about the world could, at least in theory, be said (and said more thoroughly and precisely) by microphysics.

We're now going to find that:

> (i) Botany couldn't possibly be reduced to microphysics.
>
> (ii) Psychology couldn't possibly be reduced to microphysics (or any other physical science).

And

> (iii) The reasons why psychology cannot be reduced to microphysics (or any other physical science) are distinctive. The sense in which psychology cannot be reduced to (micro)physics is more robust than the sense in which botany (or any other physical science) cannot be reduced to microphysics (or any other physical science).

Let's start by considering the classical answer to the question: "what is it to

reduce one discipline to another?" or, to put the question more precisely, "what are the conditions the fulfillment of which is both necessary and sufficient for the reduction of one discipline to another"? Here is that answer[380]:

> (CA) Let's say you're trying to reduce botany to microphysics. There are many bridges you must cross. Let *b* be an arbitrary botanical concept. So *b*, let us suppose, is the concept *pine-tree*. (A consequence of this supposition, of course, is that *x is an instance of b* is (to the same effect as) the proposition *x is a pine-tree*.) You must, at the very least, identify certain microphysical concepts $m_1 \ldots m_n$ such that any instance of the concept *pine-tree* is an instance of $m_1 \ldots m_n$. In other words, you must, to use Hempel's (1965) expression, produce an accurate "microphysical characterization" of the concept *pine-tree*. Needless to say, given *any* botanical concept, you must do the same thing *mutatis mutandis* in connection with that concept that we just did (or, rather, described doing) in connection with the concept *pine-tree*.
>
> But that is not all you must do. Let *p* be any law of botany. (If, for whatever reason, you deny that there are laws of botany, replace the word "law" with the word "principle" in what follows.) *p* is presumably to the effect that, if a botanical truth of a certain kind is true, then so (probably) is a botanical truth of a certain other kind. To reduce botany to microphysics, you must show that, whatever it is about the universe in virtue of which that general botanical truth holds there is a micro-physical characterization of it.
>
> Let us clarify this last point. Let *B* be an arbitrary botanical law. *B* demands that, if botanical concept b_1 is instantiated in space-time region *R*, then in all likelihood botanical concept b_2 be instantiated in space-time region *R**. (It is irrelevant in this context whether or not *R* and *R** are distinct regions.)
>
> Given these assumptions, this is what you must do to reduce botany to microphysics: *you must identify two sets of*

microphysical concepts, $m_1 \ldots m_k$ and $m^*_1 \ldots m^*_j$, such that (i) b_1 is instantiated in R iff $m_1 \ldots m_k$ are instantiated in R and such that (ii) b_2 is instantiated in R iff $m^*_1 \ldots m^*_j$ are instantiated in R*.*

To sum up: to reduce botany to microphysics you must provide microphysical characterizations of all botanical concepts and all microphysical laws. Doing this involves showing (1) that, for each botanical concept *b*, there are microphysical concepts $m_1 \ldots m_n$ such that the proposition *b is instantiated in R* is true if and only if the proposition $m_1 \ldots m_n$ *are (jointly) instantiated in R* is also true; and (2) that for each botanical principle *B* there is a microphysical law *M* such that an instance of *B* occurs in a given region, at a given time, only if an instance of *M* occurs in that same region, at that same time.

Of course, given *any* two disciplines D_1 and D_2, the points just made in connection with the reduction of botany to microphysics are true *mutatis mutandis* of the reduction of D_1 to D_2.

4.1 Concomitances require explanation and don’t provide it (revisited)

Right now, we’ll evaluate CA. First point: The laws governing an entity are not necessarily identical with those governing its parts. Here are some illustrations of this. *A* and *B* are two distinct battalions. General Smith is in charge of both of them. The General’s orders determine how each battalion as a whole behaves. The soldiers composing *A* are psychologically very different from those constituting *B*. So, in effect, the soldiers in *A* are not governed by the same psychological forces (laws) as the soldiers constituting *B*. Nonetheless, the forces that determine the behavior of the battalion *as a whole* coincide with those determining the other’s behavior *as a whole.* So the laws, and therefore the principles, in terms of which the behavior of *A*'s (collective) is to be understood coincide with those in terms of which *B*'s behavior is to be understood.

For the same reason *mutatis mutandis*, even if the parts composing a thing are governed by the same laws as the parts composing some other thing, it doesn't follow that, taken as wholes, those two things are governed by the same law or, therefore, are to be understood in terms of the same principles. There is considerable evidence that, even though medium-sized objects (e.g., basketballs) are *themselves* governed by deterministic laws. (More precisely, the sub-atomic parts of the atoms composing them are not governed by deterministic laws.) If this is right (and even if it isn’t, in fact), the laws governing basketballs don’t coincide with those governing their parts. (To clarify the parenthetical qualification: the laws governing basketballs coincide with those governing the behavior of objects having radically different microstructural compositions. As a result, the laws governing the micro-activity constitutive of basketballs don’t coincide with those governing other objects that, as wholes, are governed by the same laws by which basketballs, as wholes, are governed.)

Thus, two different objects could be governed by precisely the same *botanical* laws, even though the microactivity constitutive of the one is governed by entirely different laws from the microactivity of the other; and two things could be governed by different botanical laws, even though the microactivity constitutive of the one is governed by the very same laws as the microactivity constitutive of the other.

4.1.1 Complying with laws vs. being governed by them, being governed by them; being governed by them vs. being explainable in terms of them

"Your reasoning involves a serious fallacy," it will be said. "The laws of physics govern everything. Therefore, two different things cannot, at least not in the final analysis, be governed by different laws."

There are several points to make in response to this. We must start by distinguishing between "constitution" and "identity."[381] Given a statue made of clay that is still soft, one can destroy a statue (by balling up the clay) without destroying the clay. Thus, the statue and clay don't have the same properties. The one has, whereas the other lacks, the property of being destroyed when balled up. For this reason, we say that the clay *constitutes* the statue without being *identical* with it

For similar reasons, a plant is not identical with the aggregate of particles constituting it. For, by redistributing those particles (without scattering them), one can destroy the plant without destroying the aggregate. And it's probable, for the same reason *mutatis mutandis*, that mental states are not identical with the various physiological structures and patterns of physiological activity constituting them. For it's a theoretical possibility that the part of a given person's brain responsible for some psychological function is so altered that the brain-part not only continues to exist, but remains alive, even though, because of that alteration, that brain-part no longer performs *F*.

The laws of physics clearly govern and explain the behavior of the particle-aggregate *constituting* the plant. But do they govern the plant itself? No. Before we explain this answer, we must put it in context. Two microphysically identical objects are *ipso facto* botanically identical. (If neither object is a plant, they're *still* botanically identical; for neither has any botanical property not had by the other.) A corollary is that *ceteris paribus* identical microphysical changes realize identical botanical changes.

But the laws of physics don't govern plants; they govern the particle-aggregates on which the behaviors of plants supervene. In order for the laws of physics to govern plants it is necessary, but not sufficient, that they govern plant-constituting particle-aggregates. Thus must govern plants *per se*. But, as far the laws of physics are concerned, there are particles, particle-swarms, etc.---but not plants. The right position, then, is to say that the laws of physics *don't* govern plants, but that the laws of physics govern the entities that

constitute plants and that *implement* all of the changes undergone by them. Plants cannot violate the laws of physics, since a plant's condition supervenes on that of the physics-governed particle-heap of which it is composed. But those laws don't govern plants *per se.*

Here is another way of thinking about. You could coherently imagine a universe botanically identical with ours----in which plants reacted in the same way to the same stimuli and, at the cellular and super-cellular levels, were structurally just like plants in our world---but in which the laws of physics were different. In such a world, the principles in terms of which plant-behavior was to be understood would coincide with the principles in terms of which plant-behavior is actually to be understood. So the laws of botany in that world, to the extent that there is such a thing as a law of botany, would coincide with the laws of botany our world.[382] Given that the laws of physics in that world don't coincide with the laws of physics in our world, it follows that, in our world, the laws of botany—the laws that govern plants---aren't laws of physics. For exactly similar reasons, the laws of psychology in our world are not the laws physics.

So, although everything that happens to a plant is *determined* by the laws of physics, it is misleading to say that the plants themselves are governed by such laws. For laws don't govern entities whose existence they don't countenance, and the laws of physics don't countenance the existence of plants. The physicist's world is plant-free. The laws of physics govern particle-heaps on whose properties those of plants supervene but with which plants are not identical. But only the laws of botany govern such heaps *qua* plant-constituting entities. Only the laws of botany acknowledge the existence of plants. So only the laws of botany can be said to govern botanical activity. And, for the same reason *mutatis mutandis*, only the laws of psychology can be said to govern psychological activity: the laws of physics don't govern it and neither do the laws of physiology.

4.2 Concomitances require explanation and don’t provide it (revisited)

The conventional wisdom among philosophers of science is to say that, for the reasons just given, it's an *open empirical question* whether botany (or any other discipline) can be reduced to microphysics (or any other discipline)[383]:

> (CW) It cannot be known *a priori* whether or not botany is reducible to microphysics. In other words, given *only* that the microphysical laws govern the parts composing plants, it doesn’t follow that such laws govern plants themselves. So supposing that all we know about plants is that such and such laws govern the microactivity constitutive of plants, it is an open empirical question---which, in light of future findings, might or might not turn out to be answered affirmatively—whether the laws of botany coincide with, or can be derived from, the laws governing the microphysical activity composing plants.

CW is false. To say that botany can be reduced to microphysics is to say that microphysics can say whatever botany can say (and, possibly, say it better). We’ve seen that the laws governing the micro-particles composing plants *could* be different from those governing plants themselves. That means that, even if two objects are botanically indistinguishable, it could be that the one plant’s microstructure has many a property that other’s microstructure lacks. To put it another way: even if, for any botanical property phi, *x is phi* is true iff *y is phi* is true, it could be that there exists many a microphysical property psi such that *x’s microstructure has psi* is true and *y’s microstructure has psi* is false. This, in its turn, means that, for any two plants *x* and *y*, *even if* it turns out that:

> (a) for any botanical property phi, *x is phi* is true iff *y is phi* is true,

and it also turns out that:

> (b) for any microphysical property psi, *x is a psi* is true iff *y*

is a psi is true,

microphysical statements about a given plant would *still* not be analyses, or high-resolution versions, of *botanical* statements about that plant.

So even if botanical truths about plants "tracked" microphysical truths about them---even if, for any given plant, a perfect concomitance could be established between its botanical and its microphysical properties---those microphysical truths would have altogether different contents from those botanical truths. Those microphysical truths would *not* be high-resolutions of those botanical truths; those microphysical truths would *not* explicate or clarify or precisify the contents of those botanical truths.

There would thus be no reduction of botany to microphysics. The relationship of the one set of truths to the other would not be a relation of content. We would, under such circumstances, learn no more about botanical truths from microphysical truths than, were we in a world in which all round objects were green and all square ones were red, we would learn about geometrical truths from color-truths.

4.3 The sterility of brute concomitances in relation to psychophysical reduction

Bearing this in mind, let MA be a passage that is exactly like CA *except* that, in each place where CA contains a reference to botanical truths, MA contains a reference to psychology (that is otherwise just like the reference to botany that it supplanted). For reasons exactly analogous to those just stated, even if the conditions described by MA were fulfilled, then psychology would *still* not be reduced to microphysics. It's easy to extend this point. Let BA be a passage that is exactly like MA *except* that, in each place where MA contains a reference to microphysical truths, BA contains a reference to physiology (or neurology or brain-science, or whatnot). It follows for obvious reasons that psychology would *not* have been reduced to physiology (or whatnot). Thus, even if it turns out that physiological truths "track" psychological truths, the former provide not one shred of insight into the content of the latter.

Of course, should it turn that there existed some such concomitance between the physiological and the psychological, that fact could be tremendously useful. In fact, it has turned out that (with some qualifications that, although heavy, aren't relevant here) *there is* such a concomitance between the neurological and the psychological; and it has turned out that knowledge of this fact has been of tremendous use. (It has made it possible to invent psychoactive medications and, more generally, medical procedures that dramatically improve people's lives.) But, contrary to what people generally infer from this, this does *not* entail that truths about the mind are to be sought in facts about the brain. I benefit enormously from the fact that certain things happen to my computer just in case certain buttons on my computer-keyboard are struck. But the truth that such and such keys are struck on the keyboard provides not the slightest wisp of insight into the truth that, thus and such images are showing up on the monitor.

4.3.1 A variant of the point just made

Imagine a world *W* that is just like ours, except that, in it, penguins were the only animals on the South Pole and also, in that world, there were no penguins that were not on the South Pole. In such a world:

(i) *x* is an animal on the South Pole,

would be true just in case

(ii) *x* is a penguin

was also true.

Given that, in W:

(HP)[384] all and only penguins were animals on the South Pole,

it follows that, in *W*, for any given *x*, (i) and (ii) would be equivalent. But, in *W*, (i) and (ii) would mean exactly what they mean in the real world and they'd thus have very different meanings. Thus, in *W*, the meaning of:

(1) Smith is an animal on the South Pole,

would be very different from the meaning of:

(2) Smith is a penguin.

Thus, in W, the truth of (2) would not, except perhaps in some very attenuated sense, be explainable in terms of the truth of (1), or *vice versa*. Suppose that, at time *t*, Smith is the only person in place *p* and also that Smith is the only person wearing a black shirt. Under those circumstances, *Smith is in p at t* and *Smith is wearing a black shirt at t* are equivalent. But neither proposition to any degree explains the other.

The relationship between (1) and (2) is, in the relevant respects, identical with the relationship between the just-mentioned propositions about Smith.

Since, apart from the fact that HP is true in world *W* but not in our world, *W* is governed by the same laws as our world. So it is just as possible in *W* as it is in our world for something to be a penguin that is *not* on the South Pole. That is why, even in *W*, (1) doesn't explain (2), or *vice versa*; it is why, for *any x*, (i) doesn't explain (ii), or *vice versa.*

5.0 The *sui generis* nature of psychophysical explanatory gap

One of the ways in which psychology cannot be reduced to any one of the physical sciences is comparable to the way in which botany cannot be reduced to physics. But there is *a* way, we'll now argue, in which psychology cannot be reduced to any other discipline that is distinctive. The irreducibility of the psychological is distinctive; the sense in which truths of psychology are anomalous with respect to truths of any one of the physical sciences is deeper than the sense in which the truths of botany (or any other physical science) cannot be reduced to those of physics (same qualification).

5.1 Type-type vs. token-token reduction

The irreducibility of botany to microphysics lies in the fact that botanical *concepts* are not co-extensive or *a fortiori* co-intensive with microphysical *concepts.* So there is no microphysical concept that is co-extensive with the *concept* of a pine-tree. Indeed, there is no botanical concept that is co-extensive (or, therefore, co-intensive) with any micro-physical concept, that being why no microphysical principle can replace any botanical principle.

Nonetheless, given any *specific* botanical truth *B* (e.g., *plant x died on Tuesday*), it is theoretically (though not practically) possible to show *B* to be a consequence of some microphysical truth. (I am using the term "microphysical truth" to cover truths consisting of enormously long conjunctions of true microphysical propositions.) A metaphor may be of use here. If you know how the various bricks composing some building are interrelated---if you know, for any two such bricks, how they are spatially interrelated---then you can, in principle (if not in practice), figure out the shape of the structure that those bricks jointly constitute.

For exactly similar reasons, supposing for argument's sake that you knew, for some time *t*, exactly how, at t, any two of the micro-particles composing some object *O* were interrelated, you could on that basis figure out *O*'s overall structure (at least in principle). [385] Given a knowledge of *O*'s overall structure (at a given time), along with a grasp of the conditions that an object must satisfy to fall under the concept *pine-tree* or *blade of grass* or any other botanical concept, you could determine what, if any, *botanical* properties *O* had.

So even though botanical *types* can't be identified with microphysical *types*, botanical *tokens* (i.e., *instances* of botanical types) can be identified with microphysical *tokens* (i.e., *instances* of microphysical types). Moreover, even though botanical principles are not identical with, and cannot be derived from, microphysical laws, any given non-general botanical truth can (in principle, setting aside interference-effects, though not in practice) be shown to be a consequence of some non-general microphysical truth. And *this* shows that, even though the *principles* of botany cannot be derived from those of microphysics and even though, therefore, botanical *explanations* cannot be replaced by microphysical explanations, it is not anomalous that there should be botanical data in a world all of whose denizens are governed by (micro)physical laws. In other words, it shows that, even though botany

cannot, even in principle, be absorbed into microphysics, there is no reason in principle (setting aside interference-effects and other problems relating to the acquisition of knowledge of micro-particles) that, in a world all of whose *constituents* are composed of objects governed by microphysical laws and must themselves comply with such laws, there should nonetheless be *truths* that defy microphysical *explanation.*

5.2 Type-type vs. token-token reduction (continued)

But psychology is, in this respect, not to be put in the same category as botany. Not only is it not possible to reduce psychology to microphysics (or any other physical science): given any non-general biological truth, there is no microphysical truth such that---even after every possible allowance is made for interference-effects, human limitations, etc.---knowledge of that microphysical truth could possibly make it at all clear why that psychological truth should hold; there is no microphysical truth such that (even after all the usual allowances are made) makes it at all clear why any given psychological proposition should hold.

Here's why. The data modeled by any physical science---by physics, zoology, botany, etc.---is perceptual data. The data modeled by psychology is (for lack of a better word) "introspective" data. What does this mean? First of all, the "raw data" modeled by botany (for example) doesn't consist of *experiences*; it consists of things *known* (or believed to be known) through experiences. That raw data concerns plants, trees, etc. Botanists don't talk about their experiences; they talk about the things that are the objects of their experiences---the plants, etc. that they see. The same is true of physicists. They don't talk about their experiences; they talk about the things they see and touch---the needle-displacements, the discolorations on photographic-plates, etc.

By contrast, *experiences themselves* constitute the data modeled by psychology. A physicist may refer to an experience that he has ("I suddenly saw a blinding explosion and momentarily couldn't see") but only because he is attempting in so doing to describe some phenomenon known through experience (through perception); the experience is of interest only in so far as it gives one information as to some trans-experiential entity---as to something that, although the object of experience, is not itself an experience.

Psychology is interested in the experience itself, and it's interested in the object of experience only in so far as knowledge of the latter may help one acquire knowledge of the former. Psychology is interested in what I felt on such and such an occasion; in order to make it clear how I felt, I must inevitably refer to external events ("so and so took my wallet; I punched him; I was arrested; my lawyer was five hours late," etc.). In fact, descriptions of external events are constitutive of descriptions of the experiential data in which the psychologist is interested. (The contents of one's experiences are

given by “that”-clauses. “I felt *that* I had been cheated”; “I saw *that* he was foaming at the mouth.” The exceptions to this prove spurious. “I saw a dog” is really elliptical for “I saw *that* there was a dog in a certain place.” Grammatically, the object of a verb such as “see,” “feel,” or “know”---the sort of verb that would be integral to any description of one’s experience----may either be a “that”-clause *or* an ordinary noun phrase. One can say “I know *that* snow is white” or “I know Charlie.” But we have seen that, ultimately, what one believes, feels, or otherwise experiences is given by a “that”-clause.) But it is those experiences *per se*, not their objects, which the psychologist tries to model.

This is easily shown. Given two people who are atom-for-atom duplicates and whose experiences are qualitatively identical, those people are *ipso facto* identical as far as a psychologist is concerned, *even* if one of them is a brain in a vat whose experiences are caused, not by photons bouncing off of the things he thinks he sees, but by stimulations of nerve-endings, whereas the other has veridical experiences. Setting aside the irrelevant, technical issue of how the disembodied brain would make its sentiments known to anyone, the disembodied brain’s experiences are, from a psychoanalytic or psychological perspective, 100% comparable to the other person’s experiences. But to a physicist (or botanist or geologist, etc.), the fact that the one person’s experiences are hallucinatory *ipso facto* disqualifies them: they are no longer viable sources of information.

5.3 Weak vs. strong reduction and the reason why mental truths are both weakly and strongly irreducible to physical truths

Any physical science models things known through perception. Psychology does not model things known through perception. It models our experiences. Of course, our perceptions are among our experiences. But psychology studies the perceptions *themselves.* Other sciences study the things known through them. Astronomers study stars, not perceptions of stars; psychologists study perceptions of stars, not stars.

Because experiences are the subject-matter of psychology, two people whose lives are experientially indistinguishable are psychologically indistinguishable, even if one of them, during his waking life, experiences nothing but hallucinations, whereas the other has never had so much as a single hallucination.

Of course, psychologists are interested in the fact that a given person has had a hallucination—but *only* in so far as that fact indicates some psychological malfunction on that person's part. The hallucinations had by the aforementioned brain-in-a-vat are not (necessarily) so indicative, at least not necessarily, and therefore do not (necessarily) interest the psychologist *except* in so far as, and in the same way as, their veridical counterparts interest him.

5.3.1 Weak vs. strong reduction and the reason why mental truths are both weakly and strongly irreducible to physical truths (continued)

Given a building composed entirely out of bricks, if one knew how any two of those bricks were spatiotemporally interrelated, one could in principle figure out the structure of that building. Similarly, by knowing how any two atoms composing a given plant were spatiotemporally interrelated, one could in principle (though not in practice, given how many atoms would be involved) figure out what that plant's macro-characteristics were---its morphology, its mass, etc.

Further, supposing that one had such macro-knowledge and that one had a knowledge of the concepts of botany—concepts such as *plant, pine-tree, cactus, leaf*, etc.---and, therefore, that one knew what morphological, dynamic, etc. properties a thing had to have to qualify as (say) a Venus fly-trap, the aforementioned microphysical propositions would, given the propositions embodying that persons knowledge of the concepts of botany, entail (for example) that such and such was a Venus fly-trap that had thus and such distinctively botanical properties; and the same thing *mutatis mutandis* would be true of any other given botanical propositions.

So the truths of microphysics, taken together with truths describing the concepts of botany---truths that, since they merely delineate those concepts and don't say anything as to whether they're instantiated or by what, are analytic---*entail* the truths of botany. In other words, if $B_1 \ldots B_m$ is the least inclusive list containing every botanical truth, and $M_1 \ldots M_n$ and $Cb_1 \ldots Cb_o$ are the least inclusive lists containing the truths of microphysics and the truths concerning the *structures* of botanical concepts (the *analytic* truths of botany, in other words), then, given $M_1 \ldots M_n$ and $Cb_1 \ldots Cb_o$, it is 100% probable that $B_1 \ldots B_m$ are true. Put yet another way, the way in which the conjunction of $M_1 \ldots M_n$ and $Cb_1 \ldots Cb_o$ entails $B_1 \ldots B_m$ is comparable to the way in which *x is a triangle* entails *x has more than two sides.* We're dealing with deductive, not with inductive, inference. So given any *one* (non-general) botanical truth (e.g., *x is a Venus fly-trap*), there is some microphysical truth (which, in all likelihood, is a long conjunction of microphysical truths) with respect to which that botanical truth is completely non-anomalous.

5.3.2 Weak vs. strong reduction and the reason why mental truths are both weakly and strongly irreducible to physical truths (continued)

So there is *a* sense, albeit a weak one, in which botany is reducible to microphysics. Botany is reducible to microphysics in the weak sense that, for any given botanical truth, there is some microphysical truth with respect to which the former is non-anomalous. Botany is *irreducible* to microphysics in the sense that the *system* of botanical truths that botany puts forth, containing as it does botanical *principles* alongside the specific botanical truths it contains, cannot be deduced from, or otherwise understood in terms of, the truths of microphysics. In fact, we have seen that, contrary to what Hempel (1965) assumed, even if we happened to live in a world in which botanical-types happened to be co-extensive with microphysical types, botany *still* couldn't be reduced, at least not in this robust sense, to microphysics.

As we've seen, a consequence of the irreducibility, in this strong of sense, of botany as a whole to microphysics is that what is said by any given botanical proposition is different from what is said by any given microphysical proposition, even if both truths concern the same displacement of mass-energy. *Nonetheless,* any *given* botanical truth is non-anomalous with respect to some microphysical truth.

5.3.3 Weak vs. strong reduction and the reason why mental truths are both weakly and strongly irreducible to physical truths (continued)

Bearing this in mind, let's suppose that Smith has a thorough knowledge of the concepts of psychology (*desire*, *anxiety*, *intention*, etc.). Let's further suppose that Smith is given a thorough and accurate description of the microphysical events composing Brown's brain over some period of time.

By reasoning analogous to that just put forth, Smith could in principle infer everything there is to know about Brown's brain. But, given only that data, he could infer not one single psychological truth about Brown. Psychological truths (about Brown) are *not* given by propositions to the effect some external object has some property or has been displaced in a certain way. Whatever information psychological truths *do* contain, that information is *not* to the effect that brains or physiological or otherwise sense-knowable entities have (undergone) some sense-knowable property (change). That information is about the experiences *through* which such changes are known, not the changes themselves. Which, as we've said twice before, is why the psychologist *qua* psychologist regards persons *x* and *y* as psychologically indistinguishable if the propositions describing the one person's mental states coincide with those describing the other person's, it being irrelevant from his viewpoint whether there is any divergence between those two people in respect of how accurately their experiences depict the world. ("But doesn't the psychologist care whether somebody is hallucinating or not?" Yes, but only because hallucinations are typically symptoms of psychological disturbances. The envatted brain's experiences are not thus symptomatic.)

So if $M^*_1 \ldots M^*_n$ is the least inclusive list of all of the microphysical truths describing the thing constituting Brown's brain, and $P_1 \ldots P_m$ is the least inclusive list of all of the truths concerning the concepts of psychology (i.e., is the least inclusive list of all analytic psychological truths), the conjunction of those two lists of propositions does not, by itself, entail (or, indeed, otherwise confirm) so much as a single *psychological* truth about Brown.

5.3.4 Weak vs. strong reduction and the reason why mental truths are both weakly and strongly irreducible to physical truths (closing the argument)

In order for Smith to make any inferences about what is going on in Brown's mind, we must *add* to what Smith already knows. So, granting that Smith knows $M^*_1 \ldots M^*_n$ and also $P_1 \ldots P_m$, Smith would have to know that *when* thus and such microphysical propositions (about the heap of particles constituting Brown's brain) held, *then* thus and such psychological truths (about Brown) also held. In other words, in order for Smith---even if given the vast amount of information embodied in $M^*_1 \ldots M^*_n$ and $P_1 \ldots P_m$ ---to acquire knowledge of so much as a single psychological truth about Brown, he has to know of psycho-physical *correlations.*

What this shows is that the best one can do in the way of absorbing psychology into physiology is to establish correlations between them. But this means that what psychology says about the world cannot *even in principle* be deduced from, or otherwise understood in terms of, what physiology says about the world. *Even if psychological types were co-extensive with physiological types*, the very fact that psychological types are not *necessarily* coextensive with physiological types entails what psychological truths say about the world is altogether disjoint from what physiological truths say about the world.

Even in a world *W* where all red objects were round, and no non-round objects were red, the information provided by the proposition *x is red* would be different from that provided by *x is round*. For it would be true, in *W*, that, even though the *redness* and *roundness* were co-extensive types, they weren't *necessarily* so. Even in *W*, that is to say, the proposition *x is red but not round* would make sense, as would the proposition *x is round but not red*, even though they'd be false for all values of *x*. This shows that, even in a world where two properties *P* and *P** are co-extensive, unless *x is P* entails *x is P**, or *vice versa*, those propositions, if true, reveal different facts about the world.

In our world, psychological types are *not* identical with physiological types. (A given ability on somebody's part; e.g., that person's ability to drive or speak English, can migrate. It often happens that, if brain-center *A*, which thus far has mediated Smith's ability to drive, is damaged and can no longer do its job, some other brain-center *B* takes over *A*'s job---notwithstanding

that *A* and *B* are neurologically very different. There are a lot of different neurological ways to skin a given psychological cat.[386])

5.4 A similar line of thought

The truths that it is the purpose of psychology to model are such that, unless one has had some kind of *non*-perceptual encounter with those things or with things of the same kind, one couldn't possibly understand them. I obviously can't have your experiences. But unless I myself had experienced sadness, frustration, etc., I wouldn't understand what you were saying when you spoke about your own feelings of sadness, etc. (For argument's sake, we'll suppose that, so far as anyone has knowledge of what is going on in any other person's mind, I have knowledge of what is going on in your mind.) So even though I cannot possibly have the direct, intuitive knowledge of your mental states that you can have of them---even though I can do no better than to have "knowledge by description" of them, whereas you can have knowledge "by acquaintance" of them[387]---my knowledge of your mental states isn't *purely* formal; it is informed by a knowledge of relevantly similar states in myself, my knowledge of which is completely *non*-perceptual.

Moreover, the *types* of things in which the psychologist is interested are such that their instances are always known in a way that is at least *partly* non-perceptual. Their instances in the conscious or ***preconscious*** parts of one's mind are known in a *wholly* non-perceptual way.[388] Their instances in other people's minds are known in a partly non-perceptual way, as are non-(pre)conscious instances in one's own mind.

So even though the *specific* things in which the psychologist is interested are often such that he cannot know them through sense-perception, there are three fundamental differences between the subject matter of astronomy, or any other physical science, and the subject-matter of psychology:

> (i) The things in which the psychologist is interested, though known to him in a partly sense-based way (except when he is studying his own mind), are known (to the patient) in a non-sense-based way.
>
> (ii) By virtue of knowing those things in a non-sense-based way, the patient has a more robust knowledge of them than could be had by anyone who knew them in a sense-based way.

(iii) The *types* of things studied by the psychologist are such that, unless one has had direct, non-sense-based knowledge of them, one cannot even have descriptive, second-hand knowledge of their occurrence in others.

Given (i)-(iii), it follows that one's knowledge of another person's mental states is partly, though obviously not entirely, based on knowledge of one's own mental states and, therefore, that knowledge of another person's mental states is partly, though obviously not entirely, *not* sense-based.

5.4.1 Why there is a sense in which psychology is anomalous with respect to the physical that is deeper than the sense in a given physical discipline could possibly be anomalous with respect to another physical discipline

Botany, astronomy, microphysics, zoology, and geology, though very different from one another, model data to the effect that the structure of the external world is such and such; that things are spatiotemporally interrelated in such and such a way; that the things thus interrelated have such and such properties; and that those properties themselves are such that their instances are inherently constitutive, or at least dispositive, of instances of certain kinds of spatiotemporal interrelations or changes in such relations. Consider the sorts of properties in which a physical scientist is interested: temperature, conductivity, gravitational pull, acidity, malleability, density, migration patterns, physical morphology, way of reproducing, way of obtaining nourishment. It is easily shown that, for any concept *C* such that *C* is one of the concepts which is the object of (non-psychological) scientific study, *x is an instance of C* holds in virtue of the way in which mass-energy is (re)distributed .

But, for any concept *C* such that *C* is one of the concepts which is the object of psychological study, *x is an instance of C* is *not* to that effect. The proposition *Sam is sad* is not to the effect that some (re)distribution of mass-energy holds. It may be---indeed, the present author firmly believes it to be---that Sam's sadness is mediated by physiological events and therefore by (re)distributions of mass-energy. But *Sam is sad* is not to the effect that anything having any motion-constitutive or motion-conducive properties is occupying, or likely to occupy, any position relative to anything else. *Sam is sad* registers a truth that is known to Sam in a thoroughly non-perceptual way and that cannot be known to anyone other than Sam except on the basis of his own experience of sadness and therefore cannot be known to anyone other than Sam in a way that is in part (though not entirely) non-sense-based.

5.5 Donald Davidson's failed attempt to demonstrate the anomalousness of the mental with respect to the physical

Like us, Donald Davidson held that psychological truths are anomalous with respect to physical truths. He presents his argument for this in his celebrated (1970) paper "Mental Events."[389] Here is that argument:

> (DA)[390] Suppose the following. *x* is some event in Smith's brain. *x* also constitutes a desire on Smith's part to become a lawyer.
>
> Question: Why does *x* constitute such a desire? Answer: Because, given the thoughts and deeds on Smith's part that *x* follows and given the thoughts and deeds on Smith's part that *x* precedes, it makes *sense* to see *x* as a desire on Smith's part to become a lawyer. It is analytic that a desire on somebody's part to become a lawyer has certain causes and also certain effects. It is analytic that, given certain circumstances, a desire to become a lawyer leads one to study for the LSAT, apply to law school, and so on. And it's analytic that, given certain circumstances, a desire to become a lawyer be the result from one's belief that lawyering is the best way to promote the interests of justice (or whatnot).
>
> Before we take the next step in our argument, we must generalize these points. Let *P* be an arbitrary physical event and let *M* be an arbitrary mental category. (An example of a mental category would be *belief that snow is white, desire to become a senator, regret at never having learned to play a musical instrument.* So for *x* to fall into the category *desire to become a senator* is for *x* to be a desire to become a senator.) If *P* is to fall into *M*, it is necessary and sufficient that, given what *P*'s owner thought and did prior to *P*'s occurrence, and given also what *P*'s owner thought and did *after* its occurrence, it *makes sense* to suppose that *P* is an instance of *M*. In other words, for *P* to fall into *M* is for *P* to be such that it can be *reasonably interpreted* as falling into *M*. To be a belief that 1 + 1 = 2 is to be interpretable as being such a belief.
>
> This must be understood aright. Supposing that *P* is a belief

> that 1 + 1 = 2, there isn't anything *other* than *P*'s being interpretable as being such a belief that makes it be such a belief. It being thus interpretable isn't a *consequence* of its being such a belief: it *is* its being such a belief. In general, there isn't anything to anything's being an instance of mental category *M* other than its being such that it can reasonably be seen as being such an instance.
>
> To move forward, let us continue to suppose that *P* is a belief that 1 + 1 = 2. Let us also suppose that *P* is a brain-event and, in addition, that *P* falls into *physical* category *C*.
>
> For *P* to fall into *C* isn't for *P* to be such that it can reasonably be interpreted as falling into *C*. It isn't in virtue of what *P*'s owner does or thinks before or after *P*'s occurrence that *P* falls into this as opposed to that physical category. If *P* falls into a given physical category, it is in virtue of its chemical structure, or some such, it being irrelevant whether so categorizing *P* has the consequence that *P*'s owner is rational or not.
>
> It follows that, even though mental events are identical with brain-events, the reasons for which a given event is mental are different from those in virtue of which it's physical. Moreover, the former reasons cannot be understood in terms of the latter. Thus, there is an explanatory gap.

Although we've seen reason to agree with Davidson that there is an explanatory gap, DA utterly fails to establish this. This is because it depends on two false assumptions.

The first is the assumption that to be a belief that 1 + 1 = 2 (or a desire to lose weight or an intention to buy a car, etc.) is to be interpretable as being such a belief (or desire or wish, etc.). In addition to being deeply counterintuitive, Davidson's "interpretivist" view of the mental is demonstrably false. According to Davidson's interpretivism, for brain-event *x*, had by person *A*, to fall into mental category *M* is for it to be the case that, if *x* is assumed to fall into *M*, then *A* is, to that extent, rational. If this were right, nobody could possibly make any erroneous inferences; nobody could ever think or act irrationally. For, by Davidson's lights, if *x* doesn't have the consequences that a belief that 1 + 1 = 2 *should* have, it *ipso facto* isn't a

belief that 1 + 1 = 2. But people make erroneous inferences from their beliefs; people aren’t always rational. (See Chapter 19 for a more thorough refutation of Davidson’s interpretivism.) Therefore, Davidson is wrong.

Also, Davidson’s argument assumes the truth of a doctrine known as "functionalism,” which, in Chapter 22, Section 7.1, we’ll find to be false.

6.0 Duly modifying our earlier claim that to explain *x* is to show *x* to be continuous with other events in its neighborhood

Sciences model data. Hopefully these models yield truth. But, since truth is known only through the models we construct, it cannot be a scientist's immediate intention to describe the truth; his immediate intention must be to construct a model. More specifically, his immediate intention is to construct a model that, although hoped to be truth-conducive, is intended to satisfy some desideratum that can be identified without already knowing the relevant truth: some model, that is to say, that is "good" in a way that can be defined independently of the concept of truth---but that, because it is good in that way, hopefully leads to the truth.

The body of data at a person's disposal isn't homogenous. It automatically differentiates itself into sub-classes of data. These sub-classes comprise propositions that share certain key concepts and that, if accepted, tend to lead one to accept other propositions that share those concepts. If the subject makes a sufficiently intelligent and concerted effort to delineate the consequences of these propositions, they can be organized into an system of propositions whose natural tendency is to become more organized and more comprehensive.

As we've said many times, a scientist's immediate intention is to model the data, not to identify the truth. Given this fact, along with the fact that to explain is to show the apparently discontinuous to be continuous, it follows that a scientist's immediate intention to establish *data-internal* continuity, *not* among in reality *per se*---not in the data-transcendent reality to which, if all goes well, his models will lead him---but in reality *in so far* as it can be described using the concepts definitive of his discipline.

So a botanist wants to avoid *botanical* discontinuities. In his effort to explain how plant X's not being watered on Monday led to its dying on Tuesday, he'll posit botanical continuities, not microphysical continuities. From his viewpoint, the propositions by which microphysical continuities are described are data-transcendent. Thus, those propositions are meaningless from a botanical perspective, since they are constructed out of concepts that are external to those out of which botanical propositions are constructed. He'll thus posit botanical continuities. True---propositions describing such botanical continuities are exponentially more low-resolution than those describing microphysical continuities. But former propositions, unlike the

latter, are data-internal and can describe continuities that help him model the relevant data.

7.0 The concept of a datum (revisited)

We began this chapter by outlining an answer to the question: "What is a datum?" We'll end this chapter by filling in that outline.

You believe many things. Some of your beliefs are true. Some of your true beliefs are cases of knowledge. Consider the smallest set containing all of those beliefs of yours that are cases of *explanatorily robust* knowledge. A given piece of knowledge is "explanatorily robust" if the proposition thereby known is explanatorily robust, and a proposition is "explanatory robust" if it leads to explanations that lead to the truth. (So "Smith is a person who just consumed large quantities of cyanide" is explanatorily robust, since it naturally leads to an explanation of the fact that Smith is now dead.[391]) Let *S** be the smallest set containing all of the propositions that you accept by virtue of having those pieces of knowledge.

*S** can be divided into sub-classes. Some of those sub-classes are not explanatorily unified. An example of such a class would be the one whose only members are the propositions *two is even*, *Neptune is a planet*, and *ketchup is not a vegetable*. But some of those sub-classes *are* explanatorily unified. An example of such a class would be the one all and only those propositions concerning numbers that you know to be true.

Any given class corresponds to a concept.[392] The class of all plants corresponds to the concept *plant.* (This concept may be identified with the property of being a plant, at least for the purposes of this discussion, and the same thing *mutatis mutandis* is true of every other concept.[393]) The class of all things that are *either* dogs *or* numbers corresponds to the concept *thing such that x is either a dog or a number.*

7.0.1 The concept of a datum: Concepts that have explanatory value vs. those that don't

The concepts in terms of which we instinctively understand the world tend to have considerable explanatory and predictive value. The world is so structured that, given that *x* is a *plant*, it follows that it will thrive under certain conditions and die under others; that *x* cannot talk; that *x* reproduces in certain ways but not others; that *x* has certain organelles but not others; and so on. Given this information about *x*, along with a modicum of knowledge concerning *x*'s physical surroundings, it can be explained why *x* died/why, despite previously being on the brink of death *x* is thriving/whether *x* will reproduce/and so on. Thus, the concept of a plant picks out an explanatorily unified class of entities.

Not all concepts are comparable in this respect to the concept *plant.* Let a dogber be something that is *either* a dog *or* a number. Even when taken together with considerable zoological *and* mathematical knowledge, one's knowledge that *x* is a dogber is useless from an explanatory viewpoint.

Of course, *given* the knowledge that *x* is a dogber that isn't a number, or that it is a dogber that isn't a dog, one can, given a modicum of knowledge about *x*'s circumstances, make many a correct statement as to its structure and (supposing that it is a dog and therefore that, unlike numbers, things will happen to it) as to what is likely to kill it, cure it, etc. But this shows that any value had by the concept *dogber* is stolen, as it were, from some virtue had by the concept *number* or the concept *dog.*

The value of the concept *plant* is not thus derivative. The fact that *x* is a plant *by itself* has many ramifications as to *x*'s structure, etc. Of course, the fact that *x* is a rose is, in some contexts (not all), richer in predictive and explanatory information than the fact that *x* is a plant. But, in and of itself, the fact that *x* is a dogber is without any insight- or prediction-conducive ramifications. By contrast, the fact that *x* is a plant is replete with such ramifications.

7.0.2 Final preliminaries

Of those beliefs of ours that are cases of knowledge, some have explanatory value, whereas others do not. (Knowledge of a truth of the form *x is a plant* has explanatory value. Knowledge of a truth of the form *x is a dogber* does not.) Let Smith be an arbitrary person, let *S* be the set whose members include all and only those propositions that:

(i) Smith *knows* to be true,

and:

(ii) have explanatory value.

S's members can be divided into different sub-classes. Some of those sub-classes are explanatorily unified; others are not. Given any one of the explanatorily unified sub-classes, there is some set of concepts such that:

(a) Each of the members of that sub-class contains at least one of those concepts,

and such that:

(b) Given any two members of that sub-class, the one bears on the other by virtue of the fact that each contains one or more of those concepts.

If *S#* is an arbitrary one of those sub-classes, *S#* can be divided into two groups of propositions: those of whose truth Smith has *direct* knowledge and those of whose truth Smith has indirect (inferential) knowledge.
[394]

7.1 “Datum” defined

A proposition P is a datum, relative to S#, if Smith’s knowledge of it is “direct” (in the just-described) sense.

In general, *D* is a datum, relative to some class *C* of propositions known by a given person to be true, if (1) *C* is explanatorily unified, (2) *D* is a member of *C*, and (3) that knowledge of *D* is direct (in the previously described sense).

7.2 "Datum" a relative term

We've just seen that the property of being a datum is inherently relational. In this section, we'll find that this property is relative in more ways than one.

Patient *Z* dies. For the attending physician, that’s a datum. (Let’s refer to this datum as “DP.") It isn’t a conjecture. It's known directly, not through inference or *a fortiori* through theory.

But for the microphysicist, *Z*’s being dead isn’t even a fact, let alone a datum. Terms like “dead” and “ill” have no microphysical meaning. So DP is a datum for medicine, but not for microphysics.

So the concept of a “datum” is a relative notion. Something is a datum relative to one discipline but not relative to some other discipline. In fact, it’s a doubly relative notion, since what is a datum today for a given discipline may come into question tomorrow.

7.2.1 "Datum" a relative term (continued)

Also, there are degrees of datumhood. It’s *more* of a datum that I am having such and such visual and tactile sensations than it is that I am now typing on a keyboard. But it is, surely, a datum that I am typing on a keyboard. Indeed, it is much *more* of a datum than are the so-called data on which many scientific theories are based. For those data often presuppose the accuracy of instruments that, unless highly elaborate theories are correct, are quite erroneous.

The right reaction to this last fact isn’t to say that those so-called data aren’t *really* data. If one used the term "datum" *only* to refer to things known in a thoroughly theory-innocent manner, that term would refer to nothing. The right reaction is to say what we've been saying---that the concept of a datum is a relative one.

To sum up: whatever properties a piece of information must have to qualify as a datum, some pieces of information have it to a greater degree than others. At the same time, most, if not all, of those things that we describe as "data" fail to have those properties to such a high degree that there is no context in which that piece of information is a non-datum. The class of things that have those properties to that degree is either empty or highly restrictive. [395]

Part VI Mind, Body, Self

Chapter 22 Different Views as to the Relation Between Mind and Body

1.0 Some important terms and concepts

Beliefs, hopes, tickles, pains, aspirations, dreads, intentions, fantasies, and longings are *mental* entities. Chairs, explosions, rocks, gusts of air, and quagmires are *physical* entities.

How is the mental related to the physical? Here are the different views:

(1) Dualism: mental entities are *not* physical entities.[396]

(2) Materialism: mental entities *are* physical entities.[397]

(3) Dual aspect theory: the mental and the physical are different aspects of the same thing.[398]

(4) Mental eliminativism: there is no such thing as the mental. There are brain-states—but no beliefs, tickles, etc.[399]

(5) Idealism: there is no such thing as the physical. There are experiences, hopes, etc.— but no chairs, rocks, etc.[400]

Right now, let us focus on (1) and (2), since they are, pretty obviously, the only two of these five views that have any real chance of being correct. Later we'll discuss why, despite their obvious falseness, people have held (3)–(5).

First of all, (2) is *not* the view that mental entities don't exist. (2) must be distinguished at all costs from (4). According to (2), perceptions, beliefs, doubts, and so on, all exist—but they are identical with brain-states. The materialist says that, just as water is H_2O, so your belief that 1 + 1 = 2 is identical with some pattern of neural stimulation (or some such). Just as the physical chemist isn't *denying* the existence of water in identifying it with H_2O, so the materialist isn't *denying* the existence of the mental in identifying it with physical entities. Although there may be *some* tendency on the part of materialists to be dismissive of certain features of the psychological realm, they don't *have* to be thus dismissive: Chomsky is a materialist, as was Freud—and those two were nothing if not great believers in the psyche.

Advocates of (2) typically (but not universally) hold that mental entities

are identical with brain-events and brain-structures. For obvious reasons, they don't usually hold that mental entities are identical with trees or rocks or anything else that is not a part of one's body (or, more specifically, one's central nervous system).

I have published some papers arguing for dualism. I am now a materialist. But I am still fundamentally open on this issue.

Here are my own views. I believe (as I just said) that materialism is correct. Second, I believe (very strongly) that *functionalism* (a very popular form of materialism: to be defined later) is wrong. Third, although I believe that materialism is correct, I don't think (paradoxical though it may sound initially) that it can ever be explained *why* the mental is identical with (some aspects of) the physical.

Right now, let us focus on some of the basic arguments for and against (1). (In the process of discussing these arguments, we will find that there are different versions of both materialism and dualism.) Of course, since (1) is the negation of (2), any argument for (1) is an argument against (2), and any argument for (2) is an argument against (1).

2.0 Two arguments for dualism

Setting aside arguments that are patently fallacious, there are, ultimately, two arguments for dualism. (The arguments considered in sections 4.0-4.5 are, it is easily seen, variants of these arguments.)

2.1 The first argument for dualism

The mental obviously has properties not had by the physical. For example: mental entities are representational; they are purposive; they have "phenomenology" (i.e., there is something it is like to have them). Since physical events obviously don't have those properties, physical events aren't mental. (This argument was given by Franz Brentano in *Psychology from an Empirical Standpoint.*)

2.1.1 This argument evaluated

Though it has force, there is a problem with this argument. The materialist can say that physical events *do* have these properties, but that this simply isn't obvious to introspection. For example, physical events have lots of properties that aren't obvious. When you look at a chair or a rock or a pool of water, your perception doesn't tell you anything about the microstructural properties of what you are seeing—but it would obviously be absurd to conclude that what you were seeing *didn't* have invisible (to the naked eye) microstructural properties. Similarly, says the materialist, when you introspect a pain or a tickle or a belief, your introspection doesn't tell you anything about the neural properties of what you are introspecting—but it would be absurd to conclude, on that basis alone, that what you were introspecting *didn't* have neural properties.

2.1.2 A counterargument

Some mental entities seem to be *appearances.* Consider your perception of (e.g.) some nearby table. Your perception is just the event of something's *appearing* to you. It doesn't make sense to say that there is more to an appearance than meets the eye. After all, what doesn't meet the eye is *ipso facto* not an appearance. Your perception of the table obviously doesn't *seem* to consist of lots of neural events. So, if it *did* consist of such events, they are beneath the surface—they don't "meet the eye" of introspection. But given what we said a moment ago, this would mean that such events *aren't* part of that perception.

2.1.3 This counterargument evaluated

I’m genuinely open as to whether this argument is a good one. I believe (for reasons to be stated) that materialism is correct. But I cannot find any errors in the just-stated counter-response (though others claim to be able to do so). If you can identify some consideration that resolves the matter, or that gives some weight to a certain (attempted) way of resolving the matter, I would be extremely interested in what you have to say.

2.2 The second argument for dualism: the explanatory gap

Consider some psychological statement, for example:

> (P) Last year I was very sad, because I completely failed in my efforts to write a best-selling novel.

(P) is obviously the kind of statement that could be true. But can the truth of (P) be understood in terms of facts about neural-firings? Can we conceive of some statement about a person's *body* that made it clear why (P) was true? Of course, we can imagine discovering that people feel sad when, and only when, certain neurons in their brains are firing. But would such a discovery really give us any insight into the nature of the dependence (if any there be) of that psychological fact on the existence of those neural states of affairs?

Let us generalize this point. There is no denying that we can discover psycho-physical *concomitances;* that is, that we can discover truths of the form *mental phenomenon m occurs when, and only when, physical phenomenon p occurs.* But there always seems to be a "gap" between the neural event and the associated psychological event. Permit me to clarify this last point.

Water is H_2O. Of course, when you look at a glass of water, you cannot *see* that it consists of H_2O molecules. But suppose that you could shrink yourself down to the size of such a molecule. In that case, you could literally *see* the molecules with which physical chemists identify water.

When you look at a brain, you see beige tissue—you don't see emotions, beliefs, etc. If you were to shrink yourself down the size of a molecule, you *still* wouldn't see emotions, beliefs, etc. In other words, you *still* wouldn't see the entities that the materialist identifies with brain-states.[401] You would see the entities with which the *physical chemist* identifies with brain-states (atoms, molecules of various kinds) but *not,* as we just said, those with which the *materialist* identifies such states. This suggests that psychophysical identifications (e.g., "pain is c-fiber stimulation") are not comparable to typical scientific identifications (e.g., "water is H_2O"); and this makes it questionable whether psychophysical identifications are correct.

Also, the fact that you could coherently imagine seeing water molecules (even though, for many reasons, our ability to see such things is limited)

seems to be bound up (in a way that isn't easy to articulate) with the fact that identifications such as "water is H_2O" are *explanatory*. The identification of water with H_2O renders explicable otherwise inexplicable facts about water (e.g., that the amount of space occupied by a frozen gram of water is larger than that occupied by a liquid gram of water). But, at an intuitive level at least, it doesn't seem that a comparable gain of insight into the nature of pain is gained as a result of the identification of pain with c-fiber stimulation. Imagine somebody who, because of some defect, didn't feel pain. If that person were to learn everything there was to know about the structure of the neural events underlying pain, he would still be ignorant of *what pain was*. [402] By the same token, somebody who has no idea what a neuron is can know exactly what pain is.

This suggests that psychophysical identifications are, in a fundamental way, not comparable to garden-variety scientific identifications. Somebody who knows what water tastes like, looks like, etc., doesn't necessarily know what water is. But somebody who knows that water consists of H_2O molecules really does know what water is—it doesn't really matter what other knowledge he has or lacks. It thus becomes very hard to sustain the idea that (A) "pain is c-fiber stimulation" is in the same category as (B) "water is H_2O." There is a clear sense in which somebody who knows (A) may not know what pain is but in which somebody who knows (B) cannot possibly *fail* to know what water is. In connection with this, there is reason to think that there is an *explanatory gap* between the mental and the physical—to believe, in other words, that there is no way to understand the mental in terms of the physical. And if this is the case, then it would seem to follow that the mental is *not* the physical.

2.2.1 This argument evaluated

Even if the mental cannot be understood in terms of the physical, it doesn't follow that the mental isn't *identical* with the physical. Even if the fact that Jerry is sad cannot be understood in terms of the fact that Jerry's brain is in such and such a condition, it doesn't follow that Jerry's sadness isn't identical with his brain's being in that condition.

We know that biological phenomena (e.g., digestion, respiration) are identical with interactions of atomic and sub-atomic parts. But there would be no way to translate a biological statement (e.g., "Smith is now breathing") into a statement about atoms bouncing off of one another. Because statements belonging to biology involve concepts (e.g., respiration) that aren't found in physics, it is to be expected that the former are not going to be capable of being translated into, or otherwise understood in terms of, the latter. Biology talk isn't physics talk, even though all biological *events* are physical events. Similarly, psychology talk isn't physics talk (or chemistry talk or biology talk) even though psychological *events* are physical events. Because psychology uses concepts not found in any physical science (e.g., chemistry or physics), it is to be expected that statements belonging to the former are not going to be capable of being translated into, or otherwise understood in terms of, statements belonging to the latter.[403]

2.2.2. A counter-response

Although this argument obviously has merit, and is widely accepted, I don't accept it. There seems to me to be a significant difference between the relationship that holds between statements belonging to physics and statements belonging to biology, on the one hand, and the relationship holding between statements belonging to psychology and statements belonging to physics (or chemistry or biology . . .) on the other. This viewpoint is defended in Chapter 21.

2.2.3 A counter-response to the counter-response

The mental *can* be understood in terms of the physical. There is no more of an explanatory gap between the mental and the physical than between the biological and the chemical. In other words, there is no explanatory gap. Right now we'll discuss why, according to many, there is no such gap.

2.3 Is there an explanatory gap?

So far as advocates of the view just described provide clear grounds for it, those grounds lie in two doctrines. Right now, we will discuss just one of them. (In a moment, we'll discuss the other.)

Some hold that to explain why one state of affairs accompanies or follows another simply to show that, taken together, those two events instantiate a general regularity.[404] To explain why, in this one instance, smoke follows fire is simply to identify some comprehensive regularity of which this particular sequence of events is an instance. If it is shown that fire is *always* followed by smoke, it has *ipso facto* been explained why, in *this* case, fire was followed by smoke. By parity of reasoning, were we to establish perfect *concomitances* between mental events and physical events, we would *ipso facto* have explained the mental in terms of the physical.

2.3.1 A counter-response to the argument given

Hempel's conception of explanation is erroneous. The basic problem is that it's too liberal: if it were right, anything could be said to be the cause of anything. The reasons for this are given in Chapter 17. But here's the basic idea.

I grab my phone, so as to pick it up and make a call; just after I do so, it rings. By coincidence, I'm getting a call at that exact moment. Hempel's theory has the obviously false consequence that my grabbing the phone is what caused it to ring. Let E_1 be the event consisting of my grabbing the phone, and let E_2 be the event consisting of the phone's ringing. Obviously E_1 will be unique in many respects—no other event will involve *exactly* the same mass-energy distributions that it does. It follows that E_1 uniquely has some property P_1. The same thing *mutatis mutandis* is true of E_2; and E_2 therefore uniquely has some property P_2. Thus, the sequence of events consisting of E_1 and E_2 is a unique instance of the true universal generalization that any instance of P_1 is immediately followed by, and adjacent to, an instance of P_2. If Hempel's analysis were right, then my grabbing the phone *would* have caused it to ring. Since it didn't, Hempel's analysis is wrong.

2.3.2 Another counter-response

Setting this aside, it isn't widely held that mind-brain concomitances (other than obviously contrived, explanatorily irrelevant ones similar to the one just described in connection with E_1 and E_2) are to be found. The reason: it is widely believed that the mental is "multiply realizable." Instances of any kind of mental category can, it is alleged, be constituted by physical events of very different kinds; and for this reason, there are no true statements of the form "events having such and such biochemical properties occur iff events having thus and such psychological properties occur."

2.3.3 How the functionalist tries to eliminate the explanatory gap

But, even though, for the reason just given, the alleged multiple realizability of mental states seems to eliminate one way of closing the explanatory gap between mind and matter, it actually opens what many (myself not included) claim to be a promising way of closing that gap. As previously noted, almost all contemporary materialists hold that mental entities *are* multiply realizable—that, for example, a belief that snow is white could be composed of any material capable of having the causal liaisons had by a belief that snow is white. For *X* to be such a belief, it is said, is not for it to be composed of this or that material, but is for it to have certain *causes and effects.* They hold, in general, that given any mental category *M*, for a given physical entity *P* to fall into *M* simply *is* for *P* to have certain causes and certain effects. This doctrine is known as ***functionalism.***

In Sections 7.1-7.1.2 of the present chapter, it will be argued that functionalism is false, the reason being that it collapses into the obviously false view that there are *no* mental entities. It goes without saying that, if it's false, functionalism does nothing to close the explanatory gap. In any case, I'd now like to put forth a positive reason to believe that there is no way to close the explanatory gap.

3.0 The reason there is an explanatory gap

For argument's sake suppose that every mental event is identical with (or realized by) some physical event. In that case, my feeling remorseful that I stole my sister's cookie is identical with some brain-event of mine. Let *M* be my feeling of remorse, and let *B* be the corresponding brain-event. Suppose that I sincerely and accurately express my remorse—taking care to discuss its origins, its exact content, etc. Let SM be this statement. And suppose that I see an accurate image of *B* on a giant monitor, and I thoroughly and accurately describe what I see. Let SB be *this* statement.

SM describes the very same *event* as SB. But the *data* of which SM is a synthesis doesn't even overlap with the data of which SB is a synthesis. Even if mental *events* are identical with physical *events*, the *data* of which mentalistic *statements* are syntheses doesn't even overlap with the *data* of which physicalistic statements are syntheses. This means that, given the data of which any mentalistic statement is a correct synthesis, it is epistemically possible that none of the data of which any given physicalistic statement is a correct synthesis even exists. Mentalistic statements don't model the same data as physicalistic statements, or *vice versa.* So even though the *facts* described by the former probably are identical with (a subset of) the *facts* described by the latter, any statement belonging to the one category leaves it completely open whether any given statement belonging to the other is correct or not.

The purpose of Chapter 18 is to show that *there is* an explanatory gap. In that chapter, we will examine Donald Davidson's attempt to establish that there is an explanatory gap; we will find Davidson's attempt to be a failure.

4.0 Descartes' arguments for dualism

Descartes provided a straightforward and, at an intuitive level, compelling argument for (1). Here it is. I can believe that *x* is a feeling (or belief or intention) without believing that *x* is a brain-state. Therefore, *x* cannot be a brain-state if *x* is a feeling (or belief or intention).

4.1 This argument evaluated

This argument is guilty of the so-called "intensional" fallacy. Remember what we said about intensional contexts: Given only that "Smith thinks that the inventor of bifocals invented bifocals" is true, it doesn't follow that "Smith thinks that the first postmaster general invented bifocals" is true, *even though* the first postmaster general is identical with the inventor of bifocals. Similarly, given only that "Smith thinks that *x* is a belief" is true, it doesn't follow that "Smith thinks that *x* is a brain-state" is true, *even if* beliefs *are* in fact brain-states.

4.2 A refinement of Descartes' arguments for dualism

Here is an improved version of Descartes' argument that, so its proponents believe, isn't vulnerable to the points just made.

I can *coherently* think:

(*) "*x* is a feeling, *but x is not a brain-state*."

In other words, (*) is not analytically false.

All necessary truth is analytic truth. In other words, if a statement is necessarily true, that is because it holds entirely in virtue of the concepts composing its meaning (i.e., it is analytic). Therefore, given that (*) is not analytically false, it follows that it isn't necessarily false and, therefore, that it *could* be true. So feelings *could* be distinct from brain-states.

Next-step: identity holds *necessarily* if it holds at all. If *x* and *y* are identical, then they are necessarily so. Proof[405]: Suppose that *x* and *y* are identical, but only contingently so. Obviously *x* cannot be contingently identical with itself. So if *x* and *y* are contingently identical, then *y* has a property that *x* lacks: *y* has the property of being contingently identical with *x*, whereas *x* does not have that property. Thus, contrary to our supposition, *x* is *not* identical with *y*, showing that it is incoherent to suppose that there is such a thing as "contingent" identity.

Thus, given that (*) *could* be true, and given that identity holds necessarily if it holds at all, it follows that feelings (and, for exactly similar reasons, beliefs and pains) are *not* brain-states.

4.3 This argument evaluated

First of all, it is easy to adapt what we said about (i) and (ii) to this argument. So this refinement of Descartes' argument is, like the first, guilty of the intensional fallacy.

But there's another reason to question that argument. According to many contemporary philosophers, not all necessity is analytic. The sentence "water is identical with H_2O," it is said, expresses a proposition that, although non-analytic, is necessarily true. There is no possible circumstance where water is anything other than H_2O. There are circumstances where liquids that, in terms of their macroscopic properties, *resemble* water but are not H_2O. But there are no circumstances where *water* is not H_2O.

Why is this? The categories in terms of which science explains the world are known as "natural kinds." Examples of natural kinds are *zebra, metal, proton, water, emotion, human being, acid, base, star,* and *penguin.* Natural kinds are generated by properties that are predictively and explanatorily fecund. Properties that don't have those virtues (e.g., person who drives a green car, edible substance on Larry's table) don't generate natural kinds. The property of being H_2O generates a natural kind. Two things that are explanatorily and predictively very different can be indistinguishable in respect of their macro-characteristics, at least within the narrow horizons within which non-experimentalists typically operate. For this reason, that the property of having such and such macro-characteristics does not determine a natural kind.

But two things that are indistinguishable in respect of their micro-characteristics (e.g., their chemical compositions) *cannot* be explanatorily and predictively different. Thus, the property of having such and such micro-characteristics generates a natural kind.[406] It can't be known a priori what chemical composition water has. But, for the reasons just given, its having that composition, whatever it turns out to be, is essential to it. Supposing that *x* is some body of water, *x*'s being water is identical with *x*'s having that composition. Since, therefore, water necessarily has that composition—and, therefore, is necessarily H_2O—and since it cannot be known a priori that it has that composition; "water is H_2O" expresses a necessary, but non-analytic truth.

4.4 Evaluating this counter-response

This is a *prima facie* very compelling line of thought, and it is almost universally accepted. For reasons that I state in Chapter 9, I do not accept it (even though I accept some parts of it; e.g., the part about how natural kinds are individuated by their microstructures or by whatever it is about them that makes them explanatorily and predictively useful). In any case, these reflections naturally bring us to Kripke's compelling argument for dualism.

4.5 Kripke's argument for dualism

All identities hold necessarily, and some identities are indeed a posteriori. Nonetheless, dualism can be established by means of an argument not entirely unlike Descartes'. Consider some (alleged) necessary, non-analytic truth; for example:

(WH) water is H_2O.

First of all, the *sentence* "water is H_2O" must be distinguished from the corresponding proposition. Once we take into account the information *on the basis of which* one determines the identity of the proposition literally meant by that sentence, it becomes *de rigueur to se* "water is H_2O" as encoding either a proposition that is synthetic but *non*-necessary or necessary but non-synthetic. I state my reasons for holding this in Chapter 9. I state my reasons for holding this. But, for argument's sake, we'll assume the truth of the orthodox and *prima facie* plausible position that WH is both synthetic and necessary.

The reason that WH is non-analytic is that something that appeared, visually and otherwise, to be just like water might not *be* water, as it might have the wrong micro-structure. In general, whenever a statement of the form $x = y$ is non-analytic, it is because objects that aren't x can *appear* just like x. But it makes no sense to say that something that *appeared* (felt) just like a pain wasn't *actually* a pain. Anything that feels just like a pain *is* a pain. "Pain is identical with brain state *X*" would be non-analytic, if it were true. Thus, since what makes it possible for it to be non-analytic that water is H_2O is that *appearing* like water is different from *being* water, and the same thing *mutatis mutandis* being true of all non-analytic necessities, it follows that "pain is identical with brain state *X*" is not true.

5.0 The arguments for materialism

There are several powerful arguments for materialism. But there is one line of thought that lies at the core of these various arguments.

5.1 The primary argument for materialism

It is obvious that mental events create, and otherwise affect, physical events; and it is obvious that physical events create, and otherwise affect, mental events. A hot iron is pressed to your skin (physical event); as a result, you feel pain (mental event). You intend to go to the store (mental event); as a result, you end up going to the store (physical event). You feel nervous (mental event); as a result, your heart races (physical event).

There is extremely strong evidence to the effect that any physical event is the result *only* of other physical events. To explain why a person's body moves a certain way, it is necessary *only* to know what the strictly physical (neural, muscular, etc.) antecedents of that event are, and it is *not* necessary to know what its psychological antecedents are. The behavior of a person's body is no more in need of "mentalistic" explanation than is the behavior of a brick. (Of course, a knowledge of a person's psychological condition is often sufficient, given knowledge of easy-to-ascertain facts about the external world, for a knowledge of what that person is going to do next. But such knowledge is not *necessary:* such predictions can, in principle, be made without having any knowledge as to the relevant person's psychology.)

Given that every fact about a person's bodily behavior can be explained in strictly physical terms, we can take one of two positions. We can say either:

> (i) The mental does not, contrary to appearances, have any effects on the body,

or:

> (ii) The mental *does* have such effects—and the mental is *identical* with the *physical* events that we know to have those effects.

Since it is very hard to believe that the mental is without effect on the physical, (i) is almost certainly false. Hence (ii) is the right view.[407]

Evaluating this argument: My own personal view is that this argument is pretty solid; that being why I am a materialist. Others have responded (or, given their positions, would respond) differently to this argument.

For example, according to *epiphenomenalism,* even though the physical

affects the mental, the mental does *not* affect the physical.

But epiphenomenalism is clearly wrong. It's an empirical datum that the mental affects the physical and *vice versa*. Should it be denied that this is a datum, or even that it's true, I would respond by showing that epiphenomenalism is at odds with a fundamental fact about the nature of causation.

Causation is bidirectional. If my hand affects the table (e.g., if my hand pushes the table), then the table affects my hand. So if, as the epiphenomenalist believes, the physical affected the mental, then the mental would affect the physical. But in that case, epiphenomenalism is false.

The epiphenomenalist might counter-respond by saying that there are exceptions to the principle of bi-directionality that I've been advocating: physical situations create, and have effects on, shadows; but shadows don't have effects on anything.

But this move is a non-starter. Mental entities, unlike shadows, really do have causal powers. Even the epiphenomenalist doesn't deny that they affect one another. Given this fact, along with the fact mental events result from physical events, it follows that some of the energy associated with physical states of affairs is transferred to them. So unless energy is implausibly assumed to vanish from the physical world, never to return, we must assume that the mental can, so to speak, spit it back into the physical world and, therefore, that the mental *can* affect the physical.[408] And, as we've seen, once it is granted the mental can affect the physical, it must be granted that the mental *is* physical.

5.1.1 Another argument for materialism

As we've seen, the mental clearly affects the physical. If mental entities are subject to physical laws, they are *ipso facto* physical. (After all, what would it *be* to be physical, if not to obey physical laws?) And if something obeys physical laws, it is *ipso facto* physical. What would it mean to say "*X* obeys the laws of physics —but it isn't physical"? *X* is physical iff *X* obeys physical laws.

Thus, if mental entities are not physical, they are not subject to physical laws. So if mental entities were non-physical, things that *didn't* fall within the scope of physical law would constantly have effects on the physical world, resulting in physical events that couldn't be explained in terms of physical laws—physical events that were, in fact, *violations* of physical law. But such violations do not occur. In any case, there is no evidence that they occur, and certainly no evidence that they occur with anything even approaching the frequency with which, if dualism were correct, they would occur.[409]

5.1.2 A counter-response

Surely we could imagine physical objects breaking physical laws. We can imagine people sprouting wings and flying—even though, given the structures of our bodies, such events would be violations of physical law.

5.1.3 A counter-counter-response

Bertrand Russell said that objects *are* sequences of events that are governed by physical law. I agree with him.[410] If Russell is right, it makes no sense to speak of a physical object that is in violation of all physical laws. So when we imagine a person sprouting wings we *are* imagining something that is subject to physical laws (albeit laws that are different from those that are actually operative). Second, so far as the objects of one's thoughts *aren't* law-governed, those thoughts involve incoherent or otherwise impoverished views as to the nature of physical law. A defense of this controversial point of view lies outside the scope of the present book.

5.2 The third argument for materialism—Fodor's argument[411]

The mental obviously affects the physical. Anything that affects the physical is *ipso facto* physical itself. What is there to be physical other than being able to displace mass-energy (or whatever it is that events consist of)?

5.2.1 Evaluating this argument

I personally find this argument compelling, and have yet to hear a cogent response to it.

6.0 The Problems with dual aspect theory, idealism, and eliminativism

Supposing that the mental and the physical are two aspects of the same thing, we must ask: what is that thing?[412] It can't be mental; for surely the mental is not an *aspect* of the mental. (My belief that snow is white is not an *aspect* of itself or of any other mental thing.) And it can't be physical; for surely the physical is not an *aspect* of the physical. (The chair is not an *aspect* of itself or of any other physical thing.) So (3) is either incoherent *or* it presupposes that there is something spatiotemporal that is not either physical or mental. But it is incoherent to suppose that there is anything spatiotemporal that is not *either* mental *or* physical. It follows that (3) is simply incoherent.

It seems to me that so-called "dual aspect" theorists are really just materialists who have misstated their view. So far as that's correct, we can forget about (3), and focus on (2). And since we've already discussed (2), we can move on to (4).

(4) is obviously a non-starter. But, for what extremely little it's worth, here's the argument behind it. Scientific hypotheses are supposed to deal with things that can be *publicly* verified. Suppose that scientist Smith insisted that he, and he alone, were able to see (or otherwise sense-perceive) the data that confirmed some hypothesis of his. Obviously Smith wouldn't be taken seriously. My pains, tickles, etc., cannot be witnessed by anyone other than myself. For this reason, some psychologists who were attempting to scientize their discipline said that such things had no place within science and therefore had no place in a scientific world-view. Given this position, it is a short step to saying that pains, tickles, etc., don't exist at all. In a word, science deals with what is objective; mental states are subjective; therefore, mental states don't exist.

There are many fallacies in this argument. One of them is that, in actuality, the existence of pains, tickles, etc., *is* easily verified. Of course, *you* cannot have *my* pains. But you can have your pains, and you can therefore confirm that pains exist.

Also, as Searle (1992) points out, that argument involves a failure to distinguish two meanings of the term "subjective."[413] Sometimes, when we describe something as "subjective," we mean that it is influenced by personal biases. (*So and so's reaction to the incident was extremely subjective—it*

wasn't impartial and logical.) But when we say that mental entities are "subjective," we mean—not that they are biased—but that they must be *had* in order to exist. There can't be orphan pains and beliefs floating around: a belief must be *somebody's* belief. Mental states thus have, as John Searle puts it, a "first-person ontology." Obviously science mustn't be "subjective" (i.e., it mustn't be biased); but it doesn't follow that it mustn't deal with the subjective (i.e., with things having a first-person ontology). In any case, it's a plain fact that mental entities exist, and eliminativism is a hideous monstrosity to which we shall give no further thought.

(5) is not a view that anyone really holds (even though a few—e.g., George Berkeley, William James, Bertrand Russell—have claimed to do so). (5) is really more of an epistemological view than it is a view about the relation between body and mind: so far as philosophers advocate (5), it is on the grounds that it is the only viable response to *skepticism about the external world.* (Skepticism is the focus of Chapter 11 and also of Sections 6.0–8.0 of Chapter 12.)

7.0 The different versions of materialism: behaviorism

There are different versions of materialism. Let us now consider them, starting with ***behaviorism***. According to the behaviorist, to believe that 1 + 1 = 2 is to only have a disposition to respond with certain overt behaviors to certain physical stimuli. If when asked "what is 1 + 1?" you say "2"—if, in general, you respond to sensory inputs in ways that are characteristic of a belief that 1 + 1 = 2—then *in virtue of that very fact*, you believe that 1 + 1 = 2. In other words, according to the behaviorist, behaving in the right way *is* believing that 1 + 1 = 2. It isn't a *consequence* of such knowledge. It *is* such knowledge.[414]

The general definition of behaviorism is this. If *M* is any mental category, *P* is an instance of *M* just in case *P* is a disposition on the part of some individual to respond with certain overt behaviors to certain sensory inputs.

Behaviorism is false. How a person reacts to a given sensory input is a function, not only of what that input is, but also of what that person's existing mental states are. If you don't speak English, an utterance of "what is 1 + 1?" will *not* cause you to say "2." It is only in people who have various pieces of background knowledge (e.g., knowledge of linguistic conventions) that such a disposition corresponds to the knowledge that 1 + 1 = 2. What we said about that particular disposition is true (*mutatis mutandis*) of any other. By itself, no disposition to engage in a certain kind of overt behavior, given certain stimuli, corresponds to any mental state: what does so is such a disposition *plus* various pre-existing mental states. So, mental states cannot be reduced to behavioral dispositions. At most, they can be reduced to behavioral dispositions *plus* various mental states. But this means that they *can't* be reduced to behavioral dispositions.

Aware of these problems with behaviorism, but still wanting to holding onto the idea that mental entities could be dispositionalized (identified with tendencies to respond with certain overt behaviors to certain inputs), many philosophers of mind responded to this by adopting *functionalism*, which we've already discussed and will now discuss further.

7.1 Functionalism

Consider the kinds of situations that characteristically give rise to a belief that snow is white (e.g., visual perceptions of white snow), and consider all the thoughts and behaviors that characteristically result from such a belief (e.g., one's forming the belief that snow is not green). According to the functionalist, for a brain-state to be a belief that snow is white just *is* for it to have the right causes and effects. So if brain-state *B* has the right causes and effects, *B* is *ipso facto* a belief that snow is white (or a desire to become president or a pain or a feeling of euphoria). According to the functionalist, *B*'s having those causes and effects isn't *evidence* of *B's* being such a belief; nor is it a *consequence* of *B's* being such a belief. It *is B*'s being such a belief.

The functionalist says that, for any mental type *M*, mental entities are *among the characteristic causes and effects of instances of that type.* So the functionalist—unlike the behaviorist—allows that among the causes of (e.g.) a belief that snow is white are *other mental entities* (e.g., perceptions of white snow); and that among the effects of such a belief are other mental entities (e.g., a belief that snow is not green).

Remember that behaviorism falsely supposes that a person who had no linguistic knowledge would react by saying "2" to an utterance of "what is 1 + 1?" and, in general, that behaviorism identifies mental states with dispositions that can be understood in *strictly* behavioral terms. Functionalism doesn't run into this problem. Unlike behaviorism, functionalism can (supposedly) accommodate the fact that among the causes and effects of a mental entity are other mental entities. (Why "supposedly"? Keep reading.)

For something is (e.g.) a belief that snow is white wholly in virtue of its causal role. *In general, for any mental category M, there is some causal role C such that x falls into M if and only if x has C.* (This statement may not be easy to follow, but there is no simpler, but equally accurate way of saying what functionalism is.)

7.1.1 Problems with functionalism

For something to be a pain is for it to *feel* a certain way—its causal role is secondary. It is true, of course, that pains are likely to have certain causes (kicks to the shins) and certain effects (screaming loudly). But for something to be a pain isn't for it to have a certain causal role; for something to be a pain is for it to feel a certain way.

Since it says otherwise, functionalism is wrong.

There is a related point. Supposing that *P* is some pain, *P*'s having a given causal role is a *consequence* of its being a pain; its being a pain is not a consequence of its having that causal role. Supposing that *P* is some pain of yours, your writhing and screaming is an *effect* of *P*'s being a pain and, contrary to what the functionalist alleges, it thus cannot be constitutive of its being a pain.

Functionalism says that *x* is a belief that snow is white iff *x* has the right causal role. That causal role involves other mental entities. Functionalists know that. So it sounds as though functionalism analyzes the mental into causal role *plus* the mental. It thus seems questionable whether functionalism really is a form of materialism.

If cogent, this criticism of functionalism is devastating, given that the whole *raison d'être* for that doctrine is that it supposedly explains the relationship between mind and body. Functionalists respond by saying that, through judicious use of a technical device known as the "Ramsey sentence," it can be shown that functionalism *does* reduce the mental to the physical, while at the same time doing justice to the fact that, given any arbitrary mental type *M*, instances of *M* may have other mental entities for their causes and effects. Here is their reasoning (this is a paraphrase, not a quotation):

> (RS)[415] We must distinguish the concept of *all* from the concept of *any*. Among the causes and/or effects of practically every given mental entity are (a) other mental entities and (b) strictly physical inputs (e.g., photons striking one's retinas) and strictly behavioral outputs (e.g., one's reaching for the glass). But even though *any given* mental entity is to be understood in terms of other mental entities, nonetheless the *totality* of a person's mental entities can be understood entirely in terms of strictly physical inputs and strictly behavioral outputs. Let MS

(short for “mental sentence”) be a single, very large sentence that says, for any given mental kind (e.g., *belief that snow is white*, *dread of visiting one’s mother in law*, etc.), what the causes and effects of instances of that kind are. MS is going to contain “constants”: it will contain expressions, such as “dread of going jogging in the morning,” that refer to kinds of mental entities. Given any such expression, replace it with a variable. The result will be an *open sentence* (an expression that contains a *free variable but is otherwise just like a sentence*). *Bind* the occurrences of that variable in that open sentence with an existential quantifier. Do the same thing *mutatis mutandis* for each occurrence of any expression that refers to a mental kind. But don’t do it for any of the expressions that refer to physical kinds. The result will be a sentence that completely describes the causes and effects of instances of each mental kind, but that does so without using any expressions that refer to anything mental. The mental will thus be completely *described* in terms of the physical. But that sentence will do justice to the fact that mental events cause other mental events to occur, and it will do justice to the precise way in which they do so.

By saying that causal role fixes content, RS makes content too brittle: *any* deviations of *any* kind between *any* two mental states *ipso facto* amount to differences in content. No two mental entities have *exactly* the same causal roles. (There isn’t anything in your mind that has *exactly* the same causal role as my belief that 2 is an even prime. There isn’t anything in any person’s mind that has exactly the same causal role as anything in anyone else’s mind.) It follows that, if RS is right, no two people can believe that 1 + 1 = 2 or that Mongolia is in Asia, etc. Since that is obviously false, so is RS.

A story may clarify this point. Both Smith and Jones believe that snow is white. Smith is a physicist who focuses on the reflective properties of white surfaces. Jones doesn’t know anything about physics. Obviously the functional role of Smith’s belief that snow is white will be very different from that of Jones’ corresponding belief. But if, as functionalism holds, a brain-state’s being a belief that snow is white just *is* its having a certain functional role, then presumably one person’s belief that snow is white

shouldn’t be able to differ, at least not *too* drastically, from another person’s belief that snow is white.

7.1.2 The main problem with functionalism

It is not objects, but *states* of affairs, that have causal properties. What shatters the window is not *the rock*—it is rather *the rock's moving with a certain velocity, in a certain direction,* etc. Similarly, if *B* is the brain-state that realizes your belief that (e.g.) water quenches thirst, it isn't *B* that causes things to happen—it is *B*'s having these or those properties. According to functionalism, *B*'s being identical with a belief that water quenches thirst is *identical* with *B*'s having certain causes and effects—it is identical with (e.g.) its causing you to drink water when thirsty. But if *B*'s being identical with that belief is *identical* with *B*'s causing you to drink water (when thirsty), then *B*'s being such a belief cannot *cause* such behavior. Cause and effect must be distinct: if *x* causes *y*, then *x* is not identical with *y* (and neither is a part of the other). So if functionalism is correct, then *the fact that you believe that water quenches thirst* cannot cause you to do things like drink water when thirsty. Thus, if functionalism is right, then your *belief* that water quenches thirst does nothing.

Given any other behavior that we know to be caused by the existence of a certain kind of mental entity, an analogous argument shows that, if functionalism is right, such an entity cannot possibly cause that behavior. Setting aside cases of linguistic incompetence and insincerity, utterances of "snow is white" are caused by, among other things, beliefs to the effect that snow is white. Functionalism says that there is some brain-state *B* such that that *B*'s being a belief that snow is white is *identical* with, among other things, *B*'s causing that person to utter those words. But if *B*'s being such a belief is *identical* with its causing that utterance, then *B*'s being such a belief cannot *cause* that utterance to occur. So functionalism *strips* the mental of any causal powers. Given that the mental *does* have causal powers, it follows that functionalism is wrong.

8.0 Conclusion of this Chapter

The main argument for materialism: Mind and body interact. Such interactions seem not to involve violations of physical law. Given these two facts, it seems likely that materialism is correct.

The main argument for dualism: Our experiences don't *seem* to be anything like brain-states or physical entities of any other sort. So, if they are brain-states, it is because we experience them as being very different from how they really are. But it makes no sense to suppose that we experience our experiences; and it therefore makes no sense to suppose that we experience them as being different from how they are. So, given that we experience them as being totally unlike physical entities, they must really *be* totally unlike physical entities.

The argument for materialism strikes me as being slight more compelling than the argument for dualism. But second argument is hard to dismiss. It seems to show that there is something distinctive about the relationship between mind and body---that the former is not to the latter what the biological is to the chemical or what the chemical is to the physical. Given these facts, I conclude that materialism is correct, but also that the truths described by psychological statements are, for some reason other, completely and utterly incapable of being understood in terms of the truths described by physicalistic statements. In Chapter 21, I tried to explain exactly why this is so.

Chapter 23 The Computational Theory of Mind: Progressing Towards a Scientific Psychology or Regressing to Pre-Scientific Animism?

1.0 Introduction

These days, a doctrine known as the "Computational Theory of Mind" (CTM) is very popular, both among scholars (especially philosophers and computer scientists) and among the laity. CTM is a form of materialism, for reasons that will become clear.

Because of its liberal and confident use of words such as "compute" and "algorithm", CTM has an impressive sheen of rigor, and one would think that this doctrine would be informed by the same intellectual progressiveness that led to the creation and subsequent development of computer science. But we will find that, ironically, CTM involves a regression to archaic, animistic forms of thought and, more specifically, that it represents the recrudescence of the primitive and obviously spurious thinking that the grumbling of thunder represents the wrath of the gods.

Speaking more philosophically and less psychoanalytically, we'll find that CTM involves some basic ratiocinative failings. For example, CTM embodies an ignorance of the fact that it is not *objects*, but *events* that have effects. These events often *involve* objects. But objects *per se* cause nothing. *The rock* didn't break the window; what did so was the *rock's striking the window*. Also, we'll find that, if CTM were true, we couldn't know it to be true and couldn't have any legitimate reason to advocate it.

1.1 What is CTM?

According to CTM, to think is to compute. Thoughts are computations. A computation is a formal inference. So CTM is the view that any instance of thought is identical with the making of a formal inference.

But what is a "formal" inference? To qualify as "formal," a case of reasoning must be identical with a reaction, on the part of some creature or other, to symbolic-inputs. But not all such reactions qualify as cases of formal reasoning: only those reactions so qualify that aren't to any degree driven by the knowledge of the *meanings* of the symbol-tokens in question and are driven entirely by a sensitivity to their *forms.* So CTM is the contention that thought consists of responses to symbol-tokens that are driven entirely by the formal properties of those tokens and therefore embody no knowledge of what those symbol-tokens mean. In Section 1.2, we'll give some examples of such responses, thereby eliminating any obscurity from the statement just made.

Why advocate CTM? Two reasons. First, CTM, if correct, solves the mind-body problem. When *you* add two numbers, you are *ipso facto* thinking. Computers presumably add numbers. (Computers presumably compute.) Therefore, computers, when adding numbers, are *ipso facto* thinking. We know how a computer's software is related to its hardware. So CTM, if correct, solves the mind-body problem. (In other words, CTM, if correct, makes it clear how brains are responsible for the mental activity that they mediate.)

Second, CTM, if correct, solves the "tracking problem" (or so CTM-advocates say and so we'll momentarily assume). Brain-activity, though clearly 100% physics-driven, is also logic-driven (or, in any case, logic-*consistent*), at least to the extent that it mediates rational thought. So, to the extent that it's constitutive of rational thought, brain-activity, though governed by *de*scriptive laws, embodies sensitivity to *pre*scriptive laws: where such activity is concerned, the descriptive "tracks" the prescriptive. Nothing that is obeying the laws of physics is, in virtue of that fact, doing anything that can be *evaluated*--- that can be described as "correct" or "valid" or "false." (In moving hither as opposed to thither, the billiard ball isn't doing anything right or wrong, valid or invalid.) And, it seems to follow that nothing that can be so described could be constituted by brain-activity. But brain-activity *does* constitute activity that can be so described.

CTM (supposedly) solves this problem. It's possible to create an environment in which there occur symbol-tokens whose logical properties are hewed to their physical properties in such a way that, within that environment, symbol-tokens occur in sequences that are consistent with the requirements of logic. In fact, not only is it possible to create such environments: we routinely *do* create them—and employ them. We call them "computers." So it would seem that CTM, if correct, explains how activity that falls within the scope of physics can, at the same time, embody a sensitivity to the demands of logic and, more generally, of normativity.

1.2 Mentalese: the “language of thought”

According to CTM-advocates, those symbol-tokens mediating our thoughts are held to belong to a hypothetical language. This hypothetical language is generally referred to as “Mentalese.”

According to CTM, one doesn’t have to *learn* Mentalese. Brains are computers. Computers don’t function until they’ve been programmed. Therefore, our brains don’t function until they’ve been programmed. Mentalese is that program. Thus, Mentalese must be programmed into a brain *before* that brain can mediate cognitive activity, and therefore *before* that brain’s owner can learn a language or, indeed, do anything that involves having any awareness of anything. Mentalese is to a functioning brain what a computer’s operating system is to that computer. Thus, just as your IBM came “bundled” with DOS, so your brain came “bundled” with Mentalese.

In light of what we’ve said, it might seem reasonable to say that, according to CTM, one innately *knows* Mentalese. But if CTM is right, one doesn’t know Mentalese at all, let alone innately. Knowing a language involves being able to correlate linguistic with non-linguistic entities. If I couldn’t correlate snow (the powdery substance) with “snow” (the word), I couldn’t possibly know what “snow” meant. In order to correlate “snow” with snow, I must have some way of being aware of snow that doesn’t involve knowing what “snow” means. Knowing Mentalese would involve being able to correlate the expressions belonging to it with non-linguistic entities. According to CTM, thoughts are identical with Mentalese expression-tokens, and such expressions must therefore be operational if one is to make correlations or, therefore, to know a language.

So if CTM is right, a *pre*condition for one's being aware of anything is that one's brain be stocked with already operational Mentalese symbol-tokens. So it isn’t by virtue of one’s *knowing* Mentalese that expression-tokens belonging to it mediate one’s thoughts or otherwise do what CTM-advocates allege them to do. So nobody knows Mentalese and nobody understands the expressions belonging to it. Therefore, it isn’t by being understood that such expression-tokens facilitate thought.

In this respect, Mentalese is radically different from English and every other non-hypothetical language. Expression-tokens belonging to English, Urdu, etc. do not facilitate thought *except* by being understood. By contrast, the symbol-tokens of Mentalese, supposing such tokens to exist, need not be

understood to facilitate thought. Indeed, it is inherent in what Mentalese expression-tokens are that they be able to discharge *all* of their duties *without* being understood to any degree by anyone or anything.

Because Mentalese, were it to exist, would differ in this respect from English, Spanish, and other *bona fide*languages, it follows (for reasons to be discussed in Section 4.0) that the very concept of a "language of thought" is not a coherent one. Given that CTM assumes there to be a language of thought, and given that (for reasons to be disclosed in Section 4.0) no doctrine that doesn't assume this bears any resemblance to CTM, it follows that CTM is incoherent.

1.3 The concept of a formal inference: Some illustrations

We've seen, *if* correct, CTM explains much. But *is* it correct? We can't answer that unless we make it clear what, in this context, the word "compute" means. And we can't make *that* clear unless we make it clear what, in this context, it means to say of symbolic activity that it is "form-driven."

The following three stories will help answer these questions.

Story #1 (*not* an example of computation): Smith, who speaks English, is told: "Jones has two yachts." On the basis of this information, along with his knowledge of the English language, Smith concludes that the following sentence is correct: "the number *n* such that Jones has *n* yachts is such that *n* has no rational square root."

Commentary on Story #1: Smith's reasoning is cogent. But it is not in any sense "formal." Smith's thought process was a driven by an understanding of *inter alia* the *meanings* of the words he was hearing.

Story #2 (an example of *a* kind of computation): Wooden tokens of the expressions "2," "+,""9," and "=" are fed, in that order, into wood chipper *W*. Driven solely by the morphological properties of those symbol-tokens, *W* cranks out a wooden "11." And given *any* wooden-token of a sentence-fragment beginning with an operator flanked by two numerals and ending with the sign for equality, *W* does the same thing *mutatis mutandis.*

Commentary: There would seem to be *a* sense in which *W* can compute sums. It is typically *this* sense of the word "compute" that advocates of CTM have in mind when they say that to think is to compute. Why only "typically"? Why not "always"? Because sometimes what they have in mind is the *prima facie* quite different sense of the word "compute" that is illustrated in the following section.

Story #3 (an example of a different kind of computation): Smith has no idea what "1,245," "241," "1,486," etc. mean. But Smith has been told by a source whose testimony he rightly trusts that, when given ink-marks having such and such shapes, a correct sentence results when another ink-mark, having thus and such a shape, is placed in a certain position relative to the

original ink-marks. In particular, Smith knows that, given their respective shapes, the ink-marks “1,486” and “1,245 + 241 =” form a correct sentence when (some isomorph of) the former is placed to the right of (some isomorph of) the latter. It must be emphasized that Smith has no idea what any of these ink-marks mean. Smith knows nothing of relevance to this situation *other* than the rules as to how certain shapes ought to be manipulated or complemented. Those rules *in fact* describe how to do sums involving multi-digit numbers. But Smith doesn't know that. He knows only that, when given ink-marks having such and such a shape, another ink-mark must be placed beneath them if a true statement (as to whose identity he is quite ignorant) is to result.

Commentary: What Smith is doing seems to be different from what *W* is doing. Smith’s applying the just-mentioned rule involves his having various sense-perceptions (e.g., he sees various ink-deposits) and various beliefs (e.g., he believes that those ink-marks are symbols of some kind or other, and he believes that a correct statement, albeit one of whose meaning he is ignorant, will result if, underneath them, he produces an ink-mark having a certain shape); and his having such beliefs involves his making quite delicate judgments (e.g., he must judge that the ink-deposits in question were produced by sentient beings with certain intentions---he would *not* see those ink-deposits as tokening the sentence-fragment “1,245 + 241 =” were he to learn that their existence was the unintended result of some explosion involving an ink-filled urn).

Nothing comparable holds of *W*. It isn’t because of any sense-perceptions or beliefs or intentions on its part that, having been fed a wooden token of the sentence-fragment “2 + 9 =” *W* spits out a wooden “11.” So what Smith is doing in writing a “1,486” seems *not* to be comparable to what *W* is doing in spitting out a wooden “11.”

And, while there is certainly *a* sense of the word “compute” in which Smith is computing a sum, this cannot possibly be the sense in which, according to CTM, to think is to compute. A *pre*condition for Smith’s performing the just-described computation is his *already* being able to upload and process information. So CTM would be guilty of vicious circularity were it given by the thesis that, in *this* sense of the word “compute” to think is to compute. So, supposing CTM coherent, the sense in which to think is to compute is the sense in which *W* computes, and the sense in which *Smith*

computes is irrelevant.

We will now see that CTM is indefensible and that no doctrine at all similar to it can possibly be correct. There are five reasons for this. These reasons will given in Sections 2.0-6.0.

2.0 CTM's failure to distinguish between shape and syntactic form

A story will help expose one reason why CTM is false. You shout the following at me: "if your nose starts to bleed in the next ten seconds, I'll give you a million bucks." Let *T* be the sentence-token in question. As a result of *T*'s being uttered, I *do* start to bleed out of my noise, and you duly give me $1,000,000. But it wasn't because of what *T* meant that I had a nose-bleed; nor was it because of *T*'s syntax. It was *T*'s being loud that caused my nose-bleed.

My reaction here wasn't *symbolic.* Even though *T* is a symbol (or, to be precise, a symbol-token), *T*'s being a symbol did nothing. For exactly similar reasons, *W*'s spitting out an "11" isn't symbolic, even though its doing so was precipitated by events involving symbol-tokens.

2.1 This line of reasoning generalized

We've just seen that CTM is a non-starter if a symbol-token's "form" is identified with its shape. But sometimes, when referring to a symbol-token's "form," advocates of CTM are (so they say) referring to its *syntactic* form. Given this use of the word "form," CTM becomes the contention that thoughts are syntax-driven operations on symbols.

This contention is easily refuted. Syntax is semantic structure. "Smith loves Mary" and "Sally hates Greg" have the same syntax because the semantics of the one has the same *structure* as the semantics of the other.

To be sure, those sentences have very different meanings. But they're semantically isomorphic. Each constituent of the one sentence corresponds to exactly one constituent of the other ("Smith," "loves," and Mary" correspond to "Sally," "hates," and "Greg," respectively); and the relationship borne by the semantics of any given constituent of the one sentence to the semantics of that sentence as a whole coincides with the relationship borne by the semantics of that constituent's counterpart in the other sentence to the semantics of *that* sentence as a whole. So the *relation* borne by the semantics of "Smith" with respect to the proposition that Smith loves Mary coincides with the relation borne by "Sally" to the proposition that Sally hates Greg.

Given that syntax is semantic structure, syntax-driven operations are *ipso facto* meaning-driven. Being the contention that "formal" operations are *not* meaning-driven, CTM is therefore completely incoherent if taken to be the contention that "formal" operations are syntax-driven operations.

At the same time, CTM is incoherent, as we saw, if "formal" operations are taken to be shape-driven operations. There is no third way to interpret the word "formal." So CTM is incoherent.

2.2 Shape vs. syntax (revisited): The equivocation that lies at the heart of CTM

Here is the thinking embodied in CTM:

> Form-driven responses to symbolic inputs are *ipso facto* cases of thought. *W* has form-driven responses to symbolic inputs. Therefore, *W* is thinking (it is adding numbers).

But when this argument is stated in a clear and non-elliptical manner, it becomes clear that it involves an equivocation and is invalid for that reason:

> Form-driven responses to symbolic inputs are *ipso facto* cases of thought. (In other words, *syntax-driven* responses to symbol-tokens are *ipso facto* cases of thought.) *W* has form-driven responses to symbolic inputs. (In other words, W's responses to its inputs are *morphology*-driven.) Therefore *W* is thinking.

The second argument is clearly spurious. Presumably because they are aware of this, advocates of CTM typically hold that, even though a symbol-token's shape isn't *identical* with its morphology, syntax can be "coded" into morphology---that syntax can be "morphologized" in much the way that geometry can be arithmetized. This is both false and irrelevant.

It's false because anything, no matter what its shape, can have any syntactic structure, and anything, no matter what its syntactic structure, can be tokened by anything, no matter what its shape. (In other words, given an arbitrary shape, and given an arbitrary syntactic structure, there is a possible language to which there belong expressions having that syntactic structure that are tokened by objects having that shape.) It's irrelevant because, even if syntax can be morphologized, it isn't, as we've already seen, *in virtue of* its having this as opposed to that shape that a symbol-token has this as opposed to that syntax. So *in so far as* one is responding to an object's shape, one's behavior is, to that extent, *not* the effect of, or otherwise to be understood in terms of, that object's shape.

To be sure, one knows what syntax an object has *on the basis of* its shape. A large part of the reason that I know the syntax of the following sentence-

token---“John is tall and grass is green”---is that I know what shape that ink-deposit has. But sentence-tokens with precisely the same syntax can be tokened by hand-movements *or* by noises; and hand-movements and noises have no strictly morphological resemblance at all to any ink-deposit. (No ink-deposit that tokens “John is tall and grass is green” bears much resemblance to a stream of noises, or a series of hand-movements, that also tokened that same sentence, except relative to rules of projection that were contrived *ex post facto* to establish such a resemblance.)

Also, the relationship that holds between (a) my knowledge of that ink-deposit’s shape and (b) my knowledge of its syntax is of a *causal*, and not of a *constitutive*, nature. I see that it has that shape. Given that knowledge, along with my background linguistic knowledge, I judge that it has a certain syntax.

To be sure, from an experiential (phenomenological) perspective, it may seem as though *in* knowing that it has that shape I *ipso facto* know that it has that syntax. But that’s not the case. What we have is one bit of knowledge (knowledge of shape) *leading*, by way of (rapid) judgments, to some other bit of knowledge (knowledge of syntax).

Echoing what we just said, these two bits of knowledge may seem, at the level of consciousness, to come into existence simultaneously. But that appearance is misleading and, even if not misleading, irrelevant. It’s misleading because any knowledge of an object’s syntax *must* follow knowledge of its physical properties. To clarify these last two statements: If I judge that you’re angry, it’s on the basis of your facial expression, or some such. My judgment as to your emotional state may immediately follow my awareness of your expression. But given that the former is *based* on the latter and is also *caused* by it, the one awareness follows the other, however quickly. Exactly similar points hold of the relationship between my knowledge of a given object’s physical shape and my subsequent knowledge of its syntax: it isn’t possible for me to have the latter knowledge without having the former; it *is* possible for me to have the former without having the latter; and my knowledge of the former is the sole foundation for, and is also what leads to, my knowledge of the latter.

Plus, it’s irrelevant that, speaking from a strictly experiential (phenomenological) perspective, the occurrence of the one bit of knowledge *seems* to coincide with the occurrence of the other. It is the purpose, or at the very least the effect, of consciousness to condense the intricacies of the cognitions underlying it into a single, seemingly homogeneous “experience.”

If every articulation of pre-conscious cognition were faithfully represented in consciousness, if phenomenology faithfully duplicated every joint in the intricacies of the cognitions of which our conscious judgments are the outcomes, then consciousness would be so overburdened with information that it would cease to be able to perform the executive functions with which it is tasked. When seeing a rabid dog run towards me, my consciousness would present to me a richly articulated mosaic of judgments, whose consequences for my well-being would take much more time to figure out than was available to me. Fortunately, however, my consciousness presents me with a homogenous feeling that causes me to bolt into action. Instead of giving me the psychological equivalent of a two-hundred page state-department report, consciousness gives me the equivalent of an electric cattle-prod in my back. And because the rich cognitive back-story is telescoped into that unpleasant burst of phenomenology, I do the right thing. (I run to the mini-van, lock the doors, roll up the windows, and so on.) And I do it in real time.

In general, it is precisely *because* consciousness represents as homogenous "feelings" what are in fact intricate networks of judgments that the contents of consciousness, unlike their cognitive underpinnings, are *actionable.* It is, I would conjecture, the very purpose, or at least one of the purposes, of consciousness to condense cognitive intricacies into affective simplicities. Therefore, it is *inter alia* its purpose to condense distinct, albeit adjacent, stages of intricate thought processes into a single, action-causing, survival-ensuring *feeling*.

In any case, if this is at all on the mark, then it's of no relevance that my recognition of the ink-deposit's shape *seems* to be temporally coincident with my recognition of its syntax. It is to be pointed out that, if this *isn't* on the mark, then much work in psychology crumbles.

3.0 A related point: CTM incompatible with the fact that distinct thoughts can be part of some one thought-process

If CTM is right, my thinking *2 is a prime number and 2 is less than three* cannot be responsible for my later thinking *some prime number is less than three.* According to CTM, if B_1 is the totality of brain-events constitutive of my thinking *2 is a prime number and 2 is less than three*, and B_2 is the totality of brain-events constitutive of my thinking *some prime number is less than three,* it cannot be B_1's *being* such a belief that brings about B_2. So it cannot be B_1's being such a belief that leads me to believe that some prime number is less than three.

So, if CTM is right, B_1's being such a belief has nothing to do with B_2's existence or *a fortiori* with B_2's being constitutive of a thought to the effect that some prime number is less than three. Thus, my believing that two is even and is less than three cannot, if CTM is right, have anything to do with my later believing that some even prime is less than three. So, if CTM is right, T_1 has nothing to do with T_2, notwithstanding that the brain-events constitutive of T_1 brought about those constitutive of T_2.

So even though B_1 occurs right before B_2 and is causally responsible for it, and even though B_1 and B_2 are constitutive of, respectively, T_1 and T_2, nonetheless T_1 and T_2 cannot, if CTM is right, be installments in some one thought-*process.*

In general, CTM is incompatible with the fact that distinct thoughts can be part of some one thought-process. CTM is thus incompatible with the fact that thought-processes occur---with the fact that there are *sequences* of thoughts (setting aside such thought-sequences whose constituents have no thematic coherence, having at best the content-free coherence had by a series of domino-fallings).

4.0 CTM's dependence on the false assumption that syntax can mirror logical form

Context-sensitivity drives a wedge between syntax and logical form. If I say:

(1) "it is snowing *now* and it is not snowing *now*,"

what I'm saying might be true---it might stop snowing before I utter the second conjunct---even though it has the same syntax as:

(2) "*n* is even and *n* is not even."

Similarly, if I say:

(3) "that guy is identical with that guy,"

what I'm saying might be false, even though it has that same syntax as

(4) "*n* is identical with *n*."

If our thinking were syntax-driven, we couldn't recognize the differences in respect of logical form between (1) and (2) or between (3) and (4). We can recognize those differences. So our thinking is not syntax-driven. So CTM is wrong.

Here is a similar line of thought. If, on July 1, 2010, in Richmond, VA, I say:

(i) "right now, Jim is three thousand miles from here,"

the information encoded in my utterance is very different that encoded in an utterances of:

(ii) "On July 1, 2010, Jim is 3,000 miles from Richmond VA."

The information encoded in (i) doesn't follow from the information encoded in (ii), or *vice versa*. In general, when the context-sensitive expressions in a true statement are replaced with context-*in*sensitive expressions, the result is a statement that, although true, neither entails the first statement nor follows from it.

For an expression to be context-sensitive is for its *referent* to vary with the context. What an expression refers to is determined by its semantics, not its syntax. CTM says that thought is to no degree semantics-driven. If that were true, we couldn't grasp perspectival truths, such as that expressed by (i). Since we can grasp such truths, CTM is false.

5.0 CTM's reliance on the erroneous assumption that there are minimal units of thought

First a definition: a "morpheme" is an expression that cannot be decomposed into others. Bearing this in mind, let's suppose that, just as CTM says, we think in a language. In that case, our concepts are identical with expressions. Linguistic expressions consist of morphemes. (Any given complex expression; e.g., "the cat is on the mat," decomposes into expressions that cannot, in their turn, be further decomposed; (e.g., "cat," "mat"). So in contending that our thoughts are identical with (tokens of) Mentalese symbols, CTM is, in effect, assuming that we have simple concepts: concepts that cannot be decomposed in others. So CTM, if correct, entails that any given thought that one has can be decomposed into "simple" concepts. CTM thus has the radically implausible consequence that we have "simple" concepts. In other words, CTM has the consequence that some of our concepts are no more capable of being decomposed into other concepts than a morpheme (e.g., the word "cat") can be decomposed into other expressions. If this consequence of CTM is correct, it follows, as we'll now see, that some of our concepts would be such that one could have them without also having other concepts.

Let *T* be the token of the Mentalese word for "snow" that went off when, just a moment ago, I thought to myself that snow is white. We may assume that *T* is a morpheme. (If, for some reason, you reject that assumption, just map what we say about *T* onto some Mentalese expression-token that, in your judgment, is a morpheme—and you must, if you grant that we think in a language, grant that there is some such token, since morpheme-less language is a surd---and everything that we say about *T* will hold *mutatis mutandis* of this other expression-token.) Not being decomposable into other expression-tokens, *T* presumably exists independently of such tokens. Given any *simple* expression—given any morpheme—belonging to a language *L*, you can remove that morpheme without removing other morphemes. Of course, in removing that morpheme, you'll remove all of the expressions of which it is a constituent. (And, if you choose a non-lexical item; e.g., "and" or "of," you may thus do a lot of damage to *L*.) But you won't, in removing that morpheme, be removing *other* morphemes.

Supposing that CTM is right, some of our concepts must be identical with morphemes of Mentalese. Taken together with the points just made, this

entails that

> (i) If you strip Mentalese of the word-type of which *T* is a token, the *only* other expressions which, in so doing, you'll be removing from Mentalese will be other morphemes,

and also that:

> (ii) Some of our concepts are such that one can have them without having others. This means that, if CTM is right, your concept of snow is *primitive*, the same being true of your concept of Smith, Jones, and Brown (assuming, for argument's sake, that the Mentalese expressions for these individuals are morphemic).

5.1 CTM's reliance on the erroneous assumption that there are minimal units of thought (continued)

But there are no primitive concepts. I couldn't possibly have a concept of snow, or of Barack Obama, or (to take a more tendentious example) of the number three without having a network of other concepts---extension, spatiotemporal location, surface-area, sentience, etc.

Concepts are necessarily contrastive; in order to have a concept of a thing, one has to be able to distinguish that thing from other things. A prerequisite for distinguishing a given rock from a given tree is my being aware of some property had by the rock that isn't had by the tree. And a prerequisite for my having such knowledge is my *already* having concepts.

5.1.1 How some contemporary philosophers would response to this argument

Many contemporary theorists---some of them CTM-advocates, some of them not—reject this line of thought, saying that, to have a "fix," on the rock (or the tree), I need only be on the receiving end of a causal series beginning with it. Their view is that concepts are mental photographs. The relationship borne by my concept of Socrates with respect to Socrates himself is identical with the relationship borne by a photograph of Barack Obama with respect to the events (the impact of certain light-rays with Barack Obama's body) that, in due course, led to that photograph.

The position just described involves a failure to distinguish between:

(i) awareness of some entity,

and:

(ii) information that, *if known*, would *lead* to such awareness.

I see discolorations of a certain kind on a photographic plate. Given that information, along with other pieces of background information that I have, I can make various inferences. I can infer that, at some point, light bounced off of a person who was bald and was wearing a toga (and so on). So the photographic plate is information in the sense that, to a creature with the requisite background and the requisite intelligence, its existence warrants the making of an inference that will eventuate in knowledge of some entity (e.g., a person) or state of affairs (e.g., one consisting of a person wearing a toga) that might otherwise have remained unknown. (Why "*might* otherwise have been unknown," instead of "*would* otherwise have been unknown"? Because I might have been the one who took the picture and might therefore have already known of the person and the corresponding state of affairs.) The discolorations on the photographic plate are "information" in the sense that, to a suitably endowed creature, they *warrant* the making of inferences that, in due course, lead to concepts.

In any case, it is massively to CTM's discredit that it presupposes that there are representational atoms.

5.2 A similar argument

Sentences are digital structures. Setting aside cases of ambiguity, any given sentence—any given complex expression, in fact---has a unique decomposition into minimal units of significance. This means, as previously pointed out, that our thoughts decompose into minimal representational units. Which is radically implausible.

In any case, sense-perceptions don't have a digital structure, and neither, it would seem to follow, does the information that sense-perceptions encode. Given which, it would be very artificial to maintain that all post-perceptual cognition was digital or, therefore, sentential in nature.

But, for argument's sake, let's suppose that the contents of our sense-perceptions can, as some advocates of CTM maintain, be "digitized." Supposing this hypothesis correct, all the problems previously posed for CTM by perceptual content are now posed by the post-perceptual thoughts involved in this (alleged) digitization-process. If I'm translating from Albanian to English, I must speak both Albanian and English. Moreover, supposing that I'm translating competently, I have some way of evaluating the degree to which my English outputs match my Albanian inputs. If my thinking is entirely mediated by English sentences, I won't be able to do either of these things: I won't be able to translate from Albanian into English; and, if *per impossibile* I were able to do so, I couldn't verify the accuracy of my translations. By analogous reasoning, if the aforementioned post-perceptual thoughts were themselves entirely digital, they would *ipso facto* be blind to the benchmarks relative to which their own translations were to be evaluated, supposing for argument's sake that they could produce such translations to begin with.

Of course, one could say that those processes are purely causal; i.e., that they don't embody any awareness of what, given the relevant perceptual inputs, the relevant digital outputs *ought* to be. But in that case, our post-perceptual thoughts are *ex hypothesi* merely the *results of*, and are in no way *justified by*, the contents of our perceptions. But that is surely too absurd a position to take seriously.

6.0 The regress-problem

If CTM is right, then:

(ARS)[416]anything that represents anything is a symbol.

If, as CTM says, our thoughts are sentence-tokens, then surely we understand those tokens. But understanding a sentence involves being able to pair it off with its meaning. This means that one must be able to compare that sentence with something that isn't a sentence, which wouldn't be possible if all of our thoughts were sentences (or sentence-tokens).

"That doesn't follow," it might be said. "Suppose that you think that snow is white, and that your thinking this consists in there being an occurrence of some Mentalese sentence. Your understanding that sentence could consist in there occurring, in your mind-brain, various sentences of some language—call it Mentalese_2---such that, given those sentence-occurrences, you can duly interpret the original (Mentalese) sentence."

Yes, but by that logic, Mentalese_3 sentences must occur in one's mind before one can understand those Mentalese_2 sentence-occurrences, and Mentalese_4 sentences must occur in one's mind before one can understand those Mentalese_3 sentence-occurrences, and so on *ad infinitum.* So it's viciously regressive to hold, as CTM does, that symbol-tokens mediate our thoughts. More precisely, it's viciously regressive *given* the reasonable assumption that we *understand* those symbols.

6.1 How advocates of CTM deal with the regress-problem

Aware of this line of thought, advocates of CTM say that we *don't* understand the symbol-tokens that mediate our own thoughts. As William Lycan (1984, p. 217) puts it, we are simply "built to use those symbols" in the "right" ways, much as computers are "built" to use the symbols (or, rather, the things which are tendentiously and, we have seen, wrongly *called* "symbols") on which it operates.

6.2 The problem with the way advocates of CTM deal with the regress-problem

But, so far as a given thing isn't understood, it isn't functioning *as* a symbol. A story, similar to one told earlier, will make this clear. On my desk, I have a wooden sculpture, which we'll call "WS," that tokens the sentence "haste makes waste." Even though WS tokens three words, all three word-tokens are carved from the same block of wood. So WS consists of one, solid object, not of a lot of letter- or word-blocks. WS is heavy; it's made of solid oak, not pine or plywood or any other cheap sort of wood. It would thus serve as an excellent cudgel. A student of mine comes into my office. He reads it, understands it, and accordingly changes his, hitherto impetuous ways. In this context, WS is functioning as a symbol. In other words, if $E_1 \ldots E_n$ are the just-mentioned effects on the student, it is WS's being a symbol (of quite a specific kind, in fact) that leads to $E_1 \ldots E_n$. Later that afternoon, a colleague, whom I detest, comes into my office. He doesn't read WS. Because of his myopia, he sees a brown blur in the place that WS occupies. Gripped by a non-tenure-conducive rage, I strike my colleague with WS. If $E_1^* \ldots E_m^*$ are the subsequent bruises and fractures suffered by my colleague, it is WS's having a certain mass, moving with a certain velocity, etc. that brings about $E_1^* \ldots . E_m^*$. WS's *being a symbol* is utterly irrelevant. Why? Because WS's being understood is essential to its functioning as a symbol.

Let us sum up. Languages *are* sets of semantic rules. Where there are no such rules, there are no languages; and where there is no knowledge of such rules, there is no linguistic activity. The idea that there is a "language of thought" is incoherent, given that it is inconsistent with these basic truths, and so is CTM, given that it assumes that there is a language of thought.

7.0 CTM self-refuting.

CTM says that we think in sentence-tokens. CTM also says that, *qua* proposition-bearing entities, those sentence-tokens *do nothing*. For reasons that we'll now see, it follows that, if CTM is correct, it cannot be *known* to be correct or even rationally accepted.

We must start by remembering that no *thing* is a cause. The rock isn't what breaks the window. What does so is the rock's striking the window with a certain amount of force. Bearing this in mind, let *B* be the brain-structure that encodes my belief that 1 + 1 = 2. Nobody denies that *B* has causal properties. But if CTM is right, *B*'s being (or constituting) a thought to the effect that 1 + 1 = 2 does nothing. The same thing *mutatis mutandis* holds of any content-bearing brain-state. If CTM is right, even if a given brain-state bears content *C*, its bearing that content is causally inert. This entails that, if a given person accepts CTM, it isn't because of the contents of any of his thoughts that he does so; and, consequently, it isn't because he knows of any reason to accept CTM. So, if CTM is correct, there is *ipso facto* no way that a person could be justified in accepting CTM (or, by obvious extensions of this reasoning, anything else). Which, while not showing that it's false, does show that one cannot coherently advocate it.

Chapter 24 Personal and Objectual Identity

1.0 Identity-related expressions ambiguous

We often make statements of the form ‹*x* is identical with *y*›. Examples are “I’m the guy in the photo,” “water is H_2O,” “Superman is Clark Kent,” “that’s the eraser I was looking for,” and so on. Statements of this form are known as “identity statements.”

Identity statements don’t always, or even typically, use the terms “identity” or “identical with.” Usually, they just use some form of the verb “to be.” So “that’s the guy who took my money” is an identity statement, even though the terms “identity” and “identical with” are nowhere to be found in it.

1.1 Identity-related expressions ambiguous: Predicative vs. non-predicative uses of such expressions

This brings us to a fact that confused many pre-20th-century thinkers. The verb "to be" is ambiguous. As just noted, it sometimes means "identical with" ("he is George" means "he is identical with George"). But sometimes it has an entirely different meaning. When you say "Smith is tall," you are not saying that Smith is *identical* with tallness; you are not making an identity statement. So oftentimes "is" (and "are," "am," etc.) doesn't denote the relation of identity and instead denotes the relation of *having* (as in having an attribute). In other words, "is" sometimes has a *predicative* meaning: it is used to predicate a property of an individual. (To predicate a property of an individual is to attribute that property to that individual.) But in this context we will focus on the concept of identity, not on the concept of attribute possession.

That said, it should be pointed out that, although it is generally clear what is meant by statements involving the term "is" (or some other form of the verb "to be"), there are cases where it is functioning in *both* ways—where it *both* indicates identity *and* attributes a property to somebody (or something). If you say "Smith is the guy who washes my windows," that seems *both* to identify Smith with a certain person *and* to attribute various properties to him (e.g., the property of a window washer). We will find that these borderline cases expose important facts about the concept of identity.

1.2 Metaphorical uses of identity-related expressions

Identity statements are often used in a loose, figurative fashion. One often says things like “George isn’t the person he used to be—having money has changed him” and “I’m not myself today.” These statements are not meant literally. Although having money may have changed George, since he now has unsavory character traits that he used not to have, George hasn’t *literally* become a person distinct from the person he used to be. It isn’t as though George’s brain was replaced with that of an alien. Similarly, when one says “I’m not myself today,” one isn’t saying that one is literally not identical with oneself—such a statement would be nonsensical. One is merely saying that, because of fatigue (or some such), one isn’t doing as well as one would like (or some such). Henceforth, we will disregard non-literal sentences like “George isn’t himself today” and will concern ourselves only with what is literally meant by “identical with,” “identity,” etc.

1.3 Numerical vs. qualitative identity

Terms such as “identical,” “identity,” etc. sometimes refer to “strict” or “numerical” identity, and sometimes refer to “qualitative” identity.

For *x* and *y* to be ***qualitatively identical*** is for them to have the very same features (or, in any case, for them to be as similar to each other as is possible given the one is not the very same entity as the other). *x* and *y* can be qualitatively identical *even if they are not literally the same thing*—even if, in other words, *x* and *y* are two things, not one. If you are cloned, you and your clone are qualitatively identical, meaning that you and your clone are very similar. But there are still two of you.

In philosophy, we are concerned with the other kind of identity, which we will henceforth refer to as ***numerical identity.*** This is the kind of identity involved in statements such as “water is identical with H_2O.” This doesn’t say that water and H_2O are merely very similar. It says that water and H_2O are but *one* thing. Other examples of statements affirming strict identity are “the square of 4 is identical with 16,” “Clark Kent is Superman,” and “light is electromagnetic radiation.”

1.3.1 Numerical identity entails qualitative identity, but not *vice versa*

If *x* and *y* are literally the same thing, then *x* resembles *y* to the highest degree possible—after all, *x* cannot, under that circumstance, have any attribute that *y* lacks. But *x* and *y* can be qualitatively identical—that is, they can resemble each other a great deal, without being numerically identical, as we noted when discussing the possibility of cloning.

1.3.2 Correction: Numerical identity *does* entail qualitative identity and *vice versa*

Strictly speaking, qualitative identity entails numerical identity. If *x* and *y* are truly *exactly* the same—if there is *no* characteristic that the one has that the other lacks—then they must be the same. For, at any given time, they must occupy the same place and must be doing the same things.

It can be proven formally that qualitative identity entails numerical identity. If *x* and *y* are numerically identical, then, by hypothesis, neither has any property lacked by the other. *x* has the property of being identical with *x*. Therefore, *y* has the property of being identical with *x* as well. So $y = x$.

The principle that genuine qualitative identity entails numerical identity is known as the ***Identity of Indiscernibles*** (*if x and y have the very same properties, then x is numerically identical with y*.) This principle is to be contrasted with the Indiscernibility of Identicals, better known as ***Leibniz's Law*** (*if x is numerically identical with y, then x and y have the very same properties*).

A few remarks on these two principles are in order. Though it has been disputed, the Identity of Indiscernibles is clearly correct: if *x* and *y* have the same properties, then, given that *x* has the property of being identical with *x*, *y* also has that property. And, though it is seldom disputed, Leibniz's Law is not correct. When I was three, I weighed less than 100 pounds. I now weigh more than 100 pounds. So that three year old lacks some property that I have. But there is surely a sense in which I am that very person. Wrongly thinking that, if Leibniz's Law were jettisoned, logic as we know it would crumble, philosophers have responded to facts such as this one by proposing highly artificial and, more importantly, demonstrably false ways of holding onto Leibniz's Law. We'll discuss these misguided efforts later in this chapter, and we'll propose an alternative to them.

2.0 Objectual identity

Sometimes identity statements concern *people* or other animate beings, and sometimes they concern inanimate objects. Examples of the first kind are "Clark Kent is Superman" and "Einstein is the physicist who invented Relativity Theory." Examples of the second kind are "that's the shirt I just bought," "the car sitting in my driveway today is the very one I drove back in the '70s," and "the liquid in your glass is the water that you poured into it a moment ago—I haven't switched it with poison."

If an identity statement concerns an inanimate object (or substance), then it affirms a case of *objectual* identity. "Objectual" is the adjective form of "object." If an identity statement concerns people or other animate beings (e.g., dogs, Martians), then it concerns *personal* identity. (The term "personal" is used even if the animate being in question is not a person—even if it is, for example, a dog.)

2.1 Objectual identity more fundamental than personal identity

My guess is that most people would find the topic of personal identity more interesting than that of objectual identity. A bit of autobiography may make it clear why I say this.

What is it for me—a middle-aged professor who rarely drives over the speed-limit—to be the same person as the guy who, 20 years ago, was out partying every night? I have changed so much in the meantime that it almost defies comprehension how *I* could ever have been some young roustabout. My musical horizons, for example, extend no further than the easy-listening channel on my radio, and I no longer have any tolerance for the cacophony of "rock music," which used to be my life's blood. I spend my evenings either working or reading. And when I'm not engaged in one of those two activities, it's probably because I am hobbling (or, more likely, using my "Rascal"—a motor-powered personal locomotion device for the decrepit) at a rate of around 1/2 mph, from my desk to the easy chair in which I watch television shows targeted for senior citizens (e.g., "Murder She Wrote"). Not a single one of these statements would have applied to the person I was when I was your age.

This is not to mention that my once supple joints are frozen with arthritis, and my once pink and healthy vital tissues are now an oxygen-starved gray, thanks to the scourges to one's health brought on by middle age. In high school I was a creditable athlete, playing on three different teams. These days, even a game of shuffleboard pushes me to my limits; and I'm even considering giving that up, since there's no Rascal-parking anywhere near the local shuffleboard court.

It thus defies comprehension how, given my current condition, I could possibly be identical with the party animal of two decades ago. This problem, and others like it, are very interesting—much more interesting than the question of why the chair I'm sitting on now is identical with the chair I was sitting on 10 minutes ago, when I first started writing this section. In other words, personal identity seems more interesting than objectual identity.

The problem is that personal identity must be understood *in terms of* objectual identity. Why is this? There are two dimensions to the concept of personal identity, the reason being that people have both minds and bodies, and statements about personal identity often therefore depend for their truth on both bodily and mental facts. By contrast, objects (chairs, rocks, etc.)

don’t have minds—they are pure body. So identity statements concerning them (in other words, objectual identity statements) depend for their truth only on physical facts. Identity statements that concern human or animal *bodies*—for example, “this is the hand that signed that piece of paper”—concern *objectual* identity. And the reason, of course, is that bodies are *objects*—like rocks and trees. They are not *subjects*, like you and me. Statements about personal identity are often concerned *both* with physical *and* with mental facts. Since they concern physical facts and objects, such statements are concerned, at least in part, with the concept of objectual identity. One must therefore understand objectual identity *before* one can understand personal identity, just as one must understand single-variable calculus before one can understand multivariable calculus.

2.2 One kind of objectual identity: diachronic objectual identity

A story will help clarify the nature of statements concerning objectual identity. At 11:00 A.M. I make myself a sandwich. But I don't eat it; I just place it on the kitchen counter. Let *X* be the object that at *11:00 A.M.* I place on the kitchen counter. At 1:00 P.M. I see my roommate, Smith, holding and apparently intending to consume the sandwich that I made. (There is no doubt that it's the same sandwich.) Let *Y* be the object that *at 1:00 P.M.* Smith is holding.

Under these circumstances, the statement:

(*) *X* is identical with *Y,*

is true.

But exactly what does this mean? *X* doesn't have the same properties as *Y*. First of all, *X* is on the kitchen counter, whereas *Y* is in Smith's hands. *X*, having just been made out of refrigerator-stored ingredients, is nice and cool. *Y,* having been sitting on the counter for two hours is warm and yucky. Given how different *X* and *Y* are, what sense can it make to say that they are *identical?*

Here we have to take a brief detour through a bit of physics and metaphysics. *X* is not a simple entity. It has various parts—it consists of two slices of bread, a leaf of lettuce, etc. But each of those parts consists, ultimately, of a vast swarm of incessant displacements of mass-energy. Given any macroscopic physical object—a rock, a bar of soap, a plant—not a single micro-constituent of that object will stay in the same place for any amount of time. Given any one of the atoms composing any macroscopic object, that atom will move (it will rotate, vibrate, and altogether change its location) if you give it so much as a trillionth of a second. And all the objects that we deal with in everyday life—our cars, our clothes, our toothbrushes—are composed of these incessant movements of micro-particles. So, really, all of those objects are collections of *events*. The *seeming* changelessness of the poster on your wall is an illusion: At any given time, that poster consists of nothing but movements and exchanges of mass and energy.

We must make two final points before we can say what (*) means. First, in some cases, events are causally related to one another, and in some cases they're not. I brushed my teeth this morning and you put on your shoes. The

one event didn't cause the other. But my brushing my teeth did cause various particles on my teeth to dissolve or change location.

When events are causally related to each other, they satisfy one of two conditions: *either* they are adjacent in both space and time *or*, if they are not adjacent, they are connected by a series of events each of whose links *is* adjacent with at least one other event in that same series. The event of my dragging the toothbrush across the bit of cereal stuck to my tooth occupied the same approximate place as the event of the bit of cereal's leaving my tooth, and the first event came right before the second. So, in this case, the cause (my moving the toothbrush over my tooth) and the effect (the bit of cereal's being dislodged from my tooth) are adjacent in space and time.

Not all causes are related to effects in such a direct way. I turn on a flashlight by pushing a button on it. The event of my pushing this button causes a light beam to travel from my flashlight to some distant planet. So one event (my pushing a certain button) causes another event (a certain spot's appearing on an otherwise dark, planetary surface). These two events are separated in space and also in time (it takes time for light to travel, and the larger the distance, the more time is needed). Nonetheless, the cause and the effect are *connected* by a series of events each of which *is* adjacent, in space and time, with another event in that same series. Let *t* be some period before the light beam reaches the planet but after it has left my flashlight. Suppose that, at time *t*, the tip of the light beam is in place *p*. The events—the swarm of photons—in place *p* at time *t* cause there to be similar photon-swarms at the place right next to *p* at the time right after *t*. So if *p** is some place right next to *p*, and *t** is some time right after *t*, then the photon-swarms in *p* at *t* cause there to be similar photon-swarms in *p** at *t**. In their turn, *those* photon-swarms—the ones in *p** at *t**—cause there to be other, similar photon-swarms in some third place (one that is right next to *p**) at some third time (one that is right after *t**). So even though my pushing the button is remote, in both space and time, from the resulting spot of light on the distant planet, those two events are connected by a *continuous* series of events each of whose installments is caused by its predecessor and causes its successor. Any such series is known as a "causal sequence." Light beams are causal sequences; but so, for the reasons given earlier, is *every* object. *X* (the cool 11:00 A.M. sandwich on the counter) is part of the same causal sequence as *Y* (the warm 1:00 P.M. sandwich in Smith's hands).

In light of these points, we can say what (*) means. (*) says that the events

that constitute *X* and *Y* are both parts of some one causal sequence. To generalize this point, many statements of the form *a is identical with b* say that some object is causally connected in a certain way with some other object.

Here’s another example. Let *B* be the person who is, at this very moment, writing this handout; and let *A* be the earlier described wild man from many years ago. The statement *A is identical with B* is true because *B* is the result of a continuous series of changes the first of which was undergone by *A*. For reasons that I give in Chapter 14, I hold that, where there is continuity, there is causality. So the sequence just described is a "causal series," meaning that any given event in it is a cause of any later event in it.

To sum up, in some cases, *x is identical with y* says that *x* and *y* are different installments in some one causal sequence. (For reasons that I give in Chapter 14, I hold that, where there is continuity, there is causality. So the sequence just described is a "causal series," meaning that any given event in it is a cause of any later event in it.)

2.3 Another kind of objectual identity: synchronic objectual identity

But sometimes *x is identical with y* has a very different meaning. Suppose that the world's most valuable diamond is the only diamond in the Hawthorn Museum. In that case, the following identity statement is true:

> (#) "The world's most valuable diamond is identical with the only diamond in the Hawthorn Museum."

But # isn't saying anything about causal sequences. This becomes clear when we replace # with the (let us suppose) true statement:

> (#t) "The most valuable diamond in the world *at time t* is identical with the only diamond that is in the Hawthorn Museum *at time t.*"

#*t* concerns a single instant. It doesn't concern a stretch of time; and it therefore doesn't concern causal sequences, since those, by definition, take up more than an instant of time. What # and #*t* are saying is that there is some one object that has two different properties. So # is saying that:

> (##) there exists a certain object that is a diamond that is more valuable than any other diamond in the world and that, at this time, is in a certain museum (the Hawthorn museum) that contains no other diamonds.

Thus, #*t* is saying that there is some property *P* and some property *P** such that some one thing uniquely has *P* and also uniquely has *P**. Many identity statements are of this kind. Consider:

> (SC) "Superman is Clark Kent."

This doesn't tell you that two different things are one. Two things can't be one. If SC said otherwise, it would not only be false but absurd. SC makes a non-absurd statement about properties. It says of the property of being a super-hero who lives in Metropolis and has a domicile on the North pole (etc.) that exactly one thing has that property; and it says that anything that

has that property also uniquely has the property of being a wimpy reporter at the Daily Bugle who is the butt of Lois Lane's derision (etc.).

Similar remarks hold of:

(WH) "Water is H_2O (or, rather, any body of water is identical with some collection of H_2O molecules)."

WH doesn't make the absurd statement that two different things—water, on the one hand, and H_2O, on the other—are the same thing. WH makes a statement about properties. One of those properties is that of being the liquid that we drink, bathe in, etc. The other is the property of having a certain chemical composition. WH says that the first property is uniquely instantiated and, in addition, that anything instantiating that property also uniquely instantiates the second.

To sum up: $a = b$ says one of two things: (i) there is some causal sequence of which *a* and *b* are different installments or (ii) there is some one thing that has a number of different properties.

3.0 Personal identity

Suppose that each of "*a*" and "*b*" is an expression referring to a person. In that case, of course, "*a* = *b*" is an identity statement—but it is one that concerns *personal* identity, and not merely objectual identity. Nonetheless, the principles in terms of which personal identity is to be understood are pretty much the same as those in terms of which objectual identity is to be understood. For even when it is a statement about personal identity, "*a* = *b*" says either (i) there is some causal sequence of which *a* and *b* are different installments or (ii) that there are properties *P* and *P** such that each is uniquely instantiated and such that nothing instantiating the one doesn't instantiate the other.

Consider the statement:

> (IB) "The inventor of bifocals is identical with the first postmaster general."

IB says that there is some *one* person (Benjamin Franklin, as it happens) who, in a world otherwise bereft of bifocal-inventors and first post-masters general, both invented bifocals *and* was a first postmaster general. (To avoid philosophically empty technicalities relating to tense-markers, we'll assume that BF is still with us.) So IB says that somebody has a number of different properties. So, in this respect, IB is SC, #*t*, and WH.

Now consider the statement:

> (OF) "The old fogy who, at this very moment is writing this chapter (which no one will ever read and whose comedic excellence will therefore go unappreciated) is identical with the guy who, many moons ago, drove from Los Angeles to New York.

OF says that one set of events—the events constituting a person who, right now, is writing about personal identity—is an installment in a causal sequence of which an earlier installment was some set of events that constituted some person's driving from Los Angeles to New York. So OF is like our earlier statement (concerning the sandwich):

(*) *X* is identical with *Y*.

Each of OF and (*) affirms that two different, non-simultaneous events (or sets of events) are installments in some one causal sequence.

So the logical structure of statements concerning personal identity is not different from the logical structure of statements of objectual identity.

3.1 Personal identity: the mentalistic conception

Human beings have both physical and mental (i.e., psychological) characteristics. I have a certain weight. This is a physical characteristic. I have certain beliefs and interests. These are psychological characteristics. Human existence thus has a dual nature.

It is vigorously debated how the mental and the physical are related to each other. A related, equally vigorous debate concerns the nature of the events that must occur if a person is to continue to exist.

A story will clarify these points. Smith is in an accident and all of his brain is destroyed, except for the parts that regulate involuntary functions (e.g., breathing, digestion, etc.). So Smith is "brain-dead." Question: Does Smith still exist? My own view is "no." *Legally* speaking, of course, Smith exists as long as his heart is beating. But legal categories are of little use when it comes to settling philosophical issues[417], and it seems to me that Smith went out of existence the moment his mind did.

Why do I have this view? Because I have what is referred to as a "mentalistic" conception of personal identity. This is the view that you are your mind. People are identical with mental entities of some kind. This doesn't mean that people aren't also identical with physical entities. It means that *if* people are identical with physical entities (e.g., with their brains or certain parts thereof), it is only because those entities are themselves identical with mental entities. If, for example, it turns out that the mental entities with which people are identical are also brain-structures or patterns of brain activity, what follows, the mentalist believes, is *not* that people are non-mental entities but that people *are* mental entities and that mental entities, in their turn, are also physical entities.

According to mentalism, the statement "I weigh 200 pounds" really means "my *body* weighs 200 pounds" and therefore doesn't really say how much *I* weigh. And given that I am identical with purely mental entities, I weigh no more than they do—which isn't very much, since they weigh nothing if they're non-physical, and next to nothing if they *are* physical (the reason being that, if they're physical, they're patterns of neural activity that have little or no mass).

3.2 An important fact about the mentalistic conception: a materialist can have a mentalistic conception of personal identity

One can believe that mental events are identical with certain kinds of *physical* events and *still* have a mentalistic conception of personal identity. So, for example, if you believe that feelings, intentions, and mental events in general are patterns of neural activity, *and* you believe that a body that doesn't house such patterns of activity *ipso facto* doesn't house a person, then you have a mentalistic conception of personal identity. So *even if* you believe that mind facts are identical with body facts, you still have a mentalistic conception of personal identity *so long as* the relevant body facts are those that mediate mental phenomena. Most contemporary mentalists fall into this category, and I myself am inclined, with some reservations (which I discuss in Chapter 22), to think that this is the right view.

4.0 Personal identity: The non-mentalistic (or physicalistic) conception

Many philosophers these days would disagree with much of what I've just said. For these days, non-mentalistic, "bodily" conceptions of personal identity are dominant. So many contemporary philosophers would say that brain-dead Smith still exists.

Why say this? There are some obvious reasons. First of all, we can truthfully say, while pointing to Smith's thought-free body "that *person* is brain-dead." It seems a simple fact that we can describe *people* as "brain-dead" or even as just plain dead. Of course, dead people don't have anything going on in their minds. But we can still correctly say, while pointing at a corpse, "that *person* is dead."

Further, the (brain-)dead person of whom we've been speaking is identical with some *body—not* a body, like yours or mine, that mediates thoughts and feelings, but just a plain old body.

Here's a rundown of the relevant facts:

> (1) Statements such as says "that person is (brain)dead" are often true.
>
> (2) Such statements must be about people. Otherwise, it would be inappropriate to use the word "person." When pointing at (brain)dead Smith, one would have to say "that *non*-person is dead."
>
> (3) The (brain)dead Smiths of this world don't have psychological characteristics. Therefore, the right conclusion is:
>
> (4) A thing can be a person without having thoughts or any other psychological attributes.

The argument just given, though accepted by many, involves a gross overestimation of what our linguistic practices can give us in the way of philosophical illumination.[418] It's true that we refer to certain (brain)dead objects as "people." But, it seems to me, that could well just be a verbal

convention. We refer to scooters of a certain kind as “rascals.” But they’re not *really* rascals. (A rascal is a mischievous young person.)

I also think that our tendency to describe (brain-)dead bodies as “people” has emotional roots. At some level, we are holding onto the idea that, deep within that ravaged body, is a mind. In any case, the present author puts no stock in arguments such as (1)–(4).

4.1 A better argument for the physicalistic view

But there is a much better argument for the physicalistic conception of personal identity. A story will help us state this argument.

Smith is a normal adult with a vigorous mind. One day he's hit by a bus. There is a complete cessation of all mental activity. Smith's stream of mental activity has been snipped. But Smith's body doesn't die. It goes into a vegetative coma for 20 years. During that time, there is no mental activity at all in what was once Smith's brain. The brain centers regulating automatic functions (e.g., respiration, digestion) work just fine, but the parts involved in mental activity are totally inactive. No thoughts, no dreams, no sensations. Nothing. After 20 years, the body wakes up, and the hitherto inactive brain centers spring back to life. The occupant of that body is obviously in complete command of his faculties. His personality traits are just like those of the pre-accident occupant of that same body, as are the values, the memories, the cognitive aptitudes, and so on. Let *X* be the person who occupied that body before the accident and let *Y* be the person who's been occupying it since the accident.

Is *X* the same person as *Y*? Is the post-coma person Smith? Or is it somebody who is psychologically a lot like Smith, but isn't *actually* him?

Surely it really is Smith. That's what we'd say, anyway. It's very hard to believe that *Y* is a brand-new person (albeit one who has the psychological characteristics ordinarily had only by adults).

This is a serious problem for the mentalistic view of personal identity. According to the mentalist, Smith *is* his mind. No mind, no Smith. Whatever his mind is exactly, it would make no sense to say "his *mind* is fine—but there's not so much as a flicker of mental activity going on in that head of his." Thus, no thoughts, no mind. So if the mentalist is right "no thoughts" means "no mind"; which means "no Smith"; which means that the minute Smith's mental stream was cut off, Smith ceased to exist.[419]

4.1.1 An unsuccessful mentalistic rejoinder

"But so what?" a mentalist might ask:

> "Smith went out of existence 20 years ago. And he came back into existence a few days ago. The same happens to everyone every day. You go to sleep. During that time, except when you're dreaming, your mind is gone, and so, therefore, are you. You wake up, and your mind is back, and so, therefore, are you. People go away. They come back. What's the problem?"

My response: Nothing that goes out of existence comes back into existence. Ever. No exceptions. Therefore, if *you* can survive interruptions of your mental existence—if *you* can awaken from a sleep, or a coma, during which there was *no mental activity at all* in your cranium—then *you* are not a stream of mental activity. And if you are not such a stream, it's a little hard to see how *you* could be your mind. So if *you* can wake up from thought-free slumber, then *you* aren't your thoughts or, therefore, your mind; and since it was your body, and your body alone, that existed uninterruptedly during sleep-time, you must be your body.

By way of anticipation, I myself reject one of the steps in this argument. Although I do think that one is one's mind, I don't think that one's mind is one's thoughts. (I believe, for reasons to be discussed, that one's mind is a structure that gives rise to thoughts but does not itself consist of them.) But right now, instead of saying why, in my view, this argument is objectionable, I want to make it clear why it is such a powerful argument.

4.1.2 An unsuccessful mentalistic rejoinder (continued)

What most people will disagree with in this argument is the assumption that *nothing that goes out of existence comes back into existence.*

"But surely that isn't right," it will be said. "What about reincarnation? Teleportation? Can't things take breaks from existence? Can't they, so to speak, go on holiday, and then come back later?"

No. A story will make this clear. Using your laser gun, you vaporize some rock. So what was once a very sturdy, solid rock has now been dispersed into various atoms. But you manage, through some miracle of technology, to locate all of these atoms, scattered throughout space though they are, and then to reassemble them into a rock that is qualitatively just like the rock you vaporized. Question: is the new rock numerically identical with the one you vaporized? In other words, is it *the very same rock?* Or, instead, is it a distinct rock that is *extremely similar* to the old one?

Answer: it's a distinct rock that's very similar to the old one. They are, apart from their spatiotemporal properties, qualitatively but not numerically identical.

Another, similar story will reinforce this conclusion. Rock *R* is vaporized —but not by a human or anything else that can act deliberately. The blind, uncaring hand of nature is what does the vaporizing. *R*'s atoms are dispersed through the cosmos. But, by some extraordinary coincidence, they all reconvene a million years later and they attach themselves to one another, thereby forming a new rock. Let *R** be this new rock. Of course, *R** consists of exactly the same particles as *R*. Moreover, *R**'s structure is exactly like *R*'s. Finally, each of the atoms composing *R** has exactly the same role within *R**'s structure that it had within *R*'s structure. So, setting aside their spatiotemporal differences, *R* and *R** are qualitatively identical. But they're not numerically identical.

4.2 Continuity a *sine qua non* for diachronic identity

Continuity is necessary for identity (i.e., for genidentity/diachronic identity). It's only a slight exaggeration to say that it *is* identity. If, on the other side of the world, atoms were to collect and form a creature exactly like you—same physical and psychological characteristics—it wouldn't be you. (You're here. That other person is over there.) Should you predecease this doppelganger, he obviously won't *become* you; he won't take over your identity any more than he did when you were alive. It's irrelevant that *qualitatively* he's just like you. He doesn't pick up where you left off. Supposing that you die at time *t*, pre-*t* you doesn't flow uninterruptedly into post-*t* him.

Let's change the story a bit. Right now, there's no doppelganger of you. You pass on (sadly). So, for a period, there's no one in existence who has exactly your characteristics. But, by some extraordinary coincidence, various atoms come together and form a creature qualitatively just like you, who we'll call "new-you." Is new-you you? No. Does it make any difference whether the atoms composing you are the ones composing new-you? No. Once you're gone, you're gone.

4.3 Why the concept of teleportation is an incoherent one

On *Star Trek,* people were routinely teleported. Any case of teleportation has three phases. Phase 1: the disintegration phase. The teleportation candidate is decomposed into small particles. Phase 2: the movement phase. Those particles are moved at or near the speed of light to another place. Phase 3: the reintegration phase. At the new place, those particles are reassembled into a being that is qualitatively just like the one that, moments earlier, was disintegrated. So the Phase 3 entity is exactly similar to the Phase 1 entity.

Is the Phase 3 entity numerically identical with the Phase 1 entity? No. Once Captain Kirk has been disintegrated, he's gone. What shows up on the Enterprise isn't him; it is, at most, a doppelganger.

4.3.1 Structure-preserving continuity, as opposed to mere continuity, as the essence of identity

My students often object to this. Here's their argument: "Teleporting Capt. Kirk involves transforming his energy into a new form, and then, at the target location, converting that same energy back into the old form."

I agree with the first part: Teleporting Capt. Kirk would indeed involve transforming his energy into a new form. But so would dropping a one million ton weight on him or dispersing his atoms across the cosmos. "Same energy" doesn't mean "same object." I've been around for three-and-a-half decades. My energy has been around a lot longer than that. Thirty five years ago that energy came to be embodied in a structure that, through a series of piecemeal, continuous, and *structure-preserving* changes eventuated in the structure that is writing this sentence.

The changes in question had to be "structure-preserving" because, as we'll see in Sections 7.0–7.5, a continuous causal process that mediates drastic structural changes may not be enough for identity. The person currently typing this sentence arose, by a continuous series of changes, from a certain tadpole-like creature. But, structurally, I'm sufficiently different from that tadpole that it would be a stretch to say that I used to *be* it. But I'm structurally enough like the blobby infant in the picture that, given my causal connectedness to him, I can say that I used to *be* him.

4.4 Once gone, always gone

Let’s map these points onto our points about teleportation. The Phase 2 atom-scatter has *no* meaningful structural resemblance to the Phase 3 human who, when these atoms are integrated, appears on the deck of the Enterprise. The degree of structural resemblance between me and the zygote that eventually turned into me is far greater than the structural resemblance between the Phase 2 atom-scatter and the Phase 3 Enterprise occupant. We thus have no choice but to say that Kirk was obliterated the moment his atoms were scattered. Whatever it is that “beams aboard” isn’t Kirk: it is, at best, a good imitation of him.

4.5 The physicalist's high-card: the ice-man

Of course, certain *aspects* of existence can be interrupted. You can temporarily shelve your career as a hockey player. But *you* cannot be shelved. "But what if I'm frozen and kept in ice for a thousand years, and then thawed out and reanimated? Would that be me?" Yes, it would (or so I think). But what that means is that, when you were on ice, *you* were still in existence. If, instead of being frozen, you had been disintegrated and, a thousand years later, the atoms composing you were reassembled into a creature just like you, we would *not* say that *you* had come back; we'd say that, although you were gone, somebody just like you popped into existence.

So, returning to our ice-man story, if post-thaw you is identical with pre-thaw you, that's because the years of frozen stillness didn't *interrupt* your existence. "But," you protest, "my mind was frozen during that time. No dreams. No thoughts. Nothing (in the way of mental activity). So, clearly, my existence *was* interrupted."

Your *mental* life was interrupted—but not your existence *tout court.* "But," you respond, "you agreed that the post-thaw person is really *me*—that the pre-thaw person and the post-thaw person are identical, just as the person typing this sentence is identical with the person who typed the sentence before it."

Yes, I do agree. "But that makes no sense, given the reasonable supposition that (i) I *am* my mind and given that (ii) my mental existence was severed." Yes—it makes no sense *given* those two assumptions. (By way of anticipation, we'll find that (ii) is false.)

But it *does* make sense given a different supposition, namely, that you are your *body.* Your body remained in existence during the freeze years. There was no interruption. To be sure, certain forms of *activity* were interrupted. Everything associated with mentation was on hold. But it's a datum that your body was around—that, in fact, it was kept more intact during that period than most living organisms are kept for a period of more than a day. But, evidently, *you* were around.

Bottom line: a sentient creature can survive interruptions of its mental existence, but not its physical existence. Interrupted existence is non-existence. So a sentient creature is identical with its body.

That is the main argument on behalf of the physicalistic theory of personal identity. And it's a good one. To refute it, the mentalist would have to show

either (a) that existence can be interrupted *or* (b) that cessation of mental *activity* isn't the same thing as cessation of mind. (a) is a non-starter; establishing it would be like trying to establish that squares could have five sides. But (b) is quite defensible.

4.6 A defense of (b): the mentalist strikes back

People have relatively stable personality traits and beliefs. Yesterday I believed that snow is white. I still believe it, and I'll believe it tomorrow. This holds of countless other beliefs of mine, as well as attitudes, aspirations, values, fears, etc.

People change, of course. The deepest changes happen in one's early years. But the change never entirely stops, even though it does become increasingly internal to established structures.

Throughout all this change, however, there is a great deal of stability in one's beliefs, attitudes, etc., and also in one's basic personality architecture. And the change itself is, in its broader outlines, continuous. It doesn't typically happen in fits and starts; it is typically more evolutionary than revolutionary. Of course, there are revolutions—sudden epiphanies or transformations. But they are internal to a larger, continuous structure whose integrity they don't undermine.

Where there is regularity, there is typically an underlying structure; and where there is irregularity, there is typically a lack of structure. A person who has a job tends to act in the same ways over and over. This is because his activities must be internal to a certain organizational framework. By the same token, a person who doesn't have a job doesn't have to behave in those ways. Other things being equal, a person without a job behaves more erratically than one who doesn't.

In light of this, let's consider some humdrum facts about my psychology. When I woke up today, I believed that 1 + 1 = 2. When I went to sleep last night, I believed it. I obviously wasn't thinking *1 + 1 = 2* the entire time I was asleep. (And even if, by some remote chance, I was, there are many other beliefs that I have that weren't coursing through my psyche during sleep. So just let my belief that 1 + 1 = 2 represent any one of those.) If that belief simply ceased to exist when, upon my falling asleep, it became inactive, it would be a miracle that, on waking up, I resumed having that belief. That would be a causal anomaly, a case of creation *ex nihilo.* That belief, therefore, must have its basis in some kind of *structure* that endures even when it isn't active.

In fact, beliefs *are* structures. Beliefs must be distinguished from the use to which they're put. At any given time, most of your beliefs are inactive. You're not drawing on them, arguing for them, reconsidering them, or

emoting about them. But during these lulls the beliefs themselves no more cease to exist than a sheriff's gun ceases to exist when he's not firing it.

Beliefs are not happenings. They are not, as philosophers say, "occurrent." A searing pain is occurrent. So is any sense-perception or act of thought, whether conscious or unconscious. But beliefs aren't processes. Nor are attitudes or intentions or aspirations.

No individual belief changes. For a person's beliefs to change is for that person to shed some old beliefs or acquire new ones. I used to be convinced that Platonism is false; now I'm convinced that it's true. But that change was a case of one belief's being obliterated and replaced by a different one; it wasn't a change that was internal to the persistence of some one belief. Whenever a belief is "modified," what's going on is that it's being replaced with another, possibly related belief. The belief itself is a stable structure; it isn't an event; it isn't an occurrent phenomenon. What *is* such a phenomenon is the process by which a belief is jettisoned and replaced with another.

The same is true of personality traits and of ***propositional attitudes***. (A propositional attitude is an attitude towards a proposition: a desire *that it will rain*, an intention *that one make a lot of money*, a fear *that one's plane will crash*, etc.) What is occurrent isn't the intention itself; it's the process of acquiring it or of dumping it.

One's beliefs, attitudes, etc., are an important part of one's mind. Since these things are structures, it immediately follows that there is *much* to one's mind that isn't a process of any kind. What about the occurrent processes that are associated with mind? The pains, the pleasures, the sense-perceptions, the waves of doubt, the conscious acts of deliberation? These, it seems to me, are mental *actions* and are thus no more constitutive of your mind than a bodily action (e.g., one's jumping over a fence) is constitutive of your body.

One's mind isn't one's stream of mentation; it is the structure that mediates this mentation. We saw reason to believe this earlier. So the slogan "no thought, no mind" is false. The right slogan is: "no thought-enabling structure, no mind."

This slogan suggests another one: "mind = thought-enabling structure (as opposed to stream of thought)." When one is in a deep sleep or even cryogenically frozen, the structures responsible for thought are inactive, but they don't cease to exist on that account. (An elevator that isn't being used still exists.)

4.7 An objection

Let's consider an objection to the argument just presented:

> But aren't those structures brain-structures? Ice-man survives his thousand-year ice coma only because those brain-structures remain intact. Those structures are obviously physical; they're part of one's body, like one's heart or gall-bladder. So your theory of personal identity (viz., one's self is one's mind and one's mind consists of the structures underlying one's thoughts) collapses into a physicalist, and therefore a non-mentalistic, analysis of personal identity.

Mentalists usually hold that mental entities are physical entities. A mentalist who holds that thought-structures are brain-structures will hold that, when those brain-structures go, so will the corresponding person, the reason being that those brain-structures are identical with that person's mind.

The physicalist, on the other hand, holds that, no matter what happens to the parts of brain mediating sentience, a given individual survives *provided* that his heart, lungs, gall-bladder, etc. are still working. His reason is that one's self is one's body and it's a fact that one's body can remain alive, and certainly exist in some form or other, well after one's mind is gone.

The difference between physicalists and mentalists isn't that the former do, whereas the latter do not, believe the minds are brains (or brain-structures). The difference is that mentalists, unlike physicalists, believe the *only* parts of the mind to be directly responsible for personal identity are the parts that are identical with one's mind.

4.8 McMullin's position: things are more basic than events

Ernan McMullin, a respected historian of science, read an earlier version of this chapter. His response, which I will quote *verbatim*, is penetrating:

> I read your discussion of Hume on personal identity. It seems right. But I have two questions. One is about your identification of "things" as "swarms of events", with a causal connection holding the events together. That seems to me a much more Humean conception than is warranted. For me, events happen to things: things are what undergird events. No things, no events! The reason that this issue seems crucial in your response to Hume is that you point out (correctly) that when someone is asleep, there is still a "stable mental structure" that persists, one that is not a "mental happening". Is that structure a swarm of events (happenings)? Is your point that these are not mental but physical (?) happenings? I would find your response to Hume more persuasive if the "stable" mental structure were taken to be a thing to which events occur.

Dr. McMullen is making two distinct points:

> (i) things are more basic than events. Things happen to things. My contention that everything is either an event or a swarm of events is incoherent.
>
> (ii) my position that events are the ultimate constituents of the spatiotemporal world is inconsistent with belief that the mind consists of stable structures, as opposed to the fleeting will o' the wisps that populate consciousness.

I sent a response to Dr. McMullin in which I responded to both questions and which, because it expresses my current views, I will quote *verbatim*:

The brain-structures that give rise to fleeting mental occurrences, and that constitute the self, do consist of swarms of events; but those structures don't consist of mental events. Consider, for example, your ability to speak English. Presumably, that ability is constituted by some brain-structure. The neurons (or whatever) that, in their turn, constitute that structure are, in *their* turn, constituted by (I'm skipping some intermediate levels of description) minute displacements of mass-energy. Those displacements are not themselves mental, even though, ultimately, they're mind-constitutive. (No contradiction here: the atoms that compose cells are not themselves biological entities, even though those cells are such entities.)

In connection with your question about mental structures, you make an important point that, if correct, undermines much of what I say but that, I will now try show, is not correct. (In what follows, I'll use the word "thing" in the narrowest possible sense. So, as I'll be using it, rocks, apples, and trees are "things," but neither non-spatiotemporal entities, e.g. possibilities, nor spatiotemporal but intangible entities, e.g. gravitational fields, qualify as "things.")

Just as you say, my position is that events are the ultimate constituents of the spatiotemporal world and, therefore, that:

> (i) Although there could be events in the absence of things, there couldn't be things in the absence of events.

Your position is that things (substances), not events, are the ultimate constituents of the spatiotemporal world and, therefore, that:

> (ii) Although there could be things in the absence of events, there couldn't be events in the absence of events.

The first question to ask is: what do you believe "things" to

be? There are three possibilities:

> (a) Things consist entirely of property-instances. Consider the rock's weight, shape, color, etc. There is nothing underlying those property instances; there are just the property instances.
>
> (b) Things are identical with entities---or, as they're traditionally called, "substances"--- that *underlie* property-instances but do not themselves consist of such instances.
>
> (c) Things consist partially of property-instances and partially of substances.

For argument's sake, let's suppose that either (b) or (c) is correct. In that case, underlying a given rock's shape, mass, color, position, etc., there is something that has no shape, mass, color, etc., i.e. there is something that, although it *underlies* properties, is itself featureless. A featureless thing is a surd. Nothing can have no properties. Thus, both (b) and (c) are false.

So (a) must correct. But if (a) is right, then events are the most basic constituents of spatiotemporal world. Here's why. First of all, for an event to occur is for some property to either come to be instantiated or to cease to be instantiated. (In this context, "property" is short for "property or multiplicity of properties.")

There would be no events if there were no changes, and there would be no changes if there were no cases of properties coming or ceasing to be instantiated. Further, there would be no passage of time if there were no events. To say that it's been a year since you've seen Smith *is* to say that certain periodic processes have happened a certain number of times since you last saw Smith. Thus, period during which absolutely nothing happens cannot possibly last for more than an instant.

It follows that any case of a property-instance's existing for

a non-null period of time is parasitic on cases of change and, therefore, on events. Since, as we've established, "things" are wholly constituted by property-instances, it follows that any case of a thing's enduring for more than instant is parasitic on the existence of change.

The concept of a thing that exists for no more than an instant is incoherent. Things necessarily consist of property-instances, as we've seen. So a thing's existing for a mere instant would involve there being some property P such that, within a time-interval of zero duration, P came to be and also ceased to be instantiated. Those events, involving as they do the very same property-instance, cannot possibly be simultaneous. So the sequence consisting of those two events would necessarily occupy a time-interval that was greater than an instant. Thus, there could not possibly be a thing that existed for just one instant. Since, as we saw a moment ago, any case of a thing's enduring for more than instant is parasitic on the existence of change, it follows that changes, or events, are the most fundamental and, ultimately, the only occupants of space-time.

That said, our views may not be as far apart as they initially seem to be. The concept of substance is incoherent *if* it's taken to coincide with the concept of a featureless entity (albeit one that underlies property-instances). But there is another way to understand the concept of substance, relative to which that concept *is* coherent. Assuming the concept of substance to be a coherent one, substances are forces of a certain kind. They are those forces that bind events together in such a way that the resulting structure has prediction/explanation-enhancing properties characteristic of rocks, statues, and other paradigm-cases of so-called "things."

But if substances are forces that bind events together into things, then substances are causal relations holding among events. So, assuming the concept of substance to be a coherent one, things are composed wholly of events. Which, taken in conjunction with some of the points made earlier, entails that events are the most fundamental constituents of the

spatiotemporal world.

In conclusion, there is no coherent defense of the position that events are *not* the most fundamental constituents of the spatiotemporal world or, consequently, of the view that any of the spatiotemporal world's occupants are not either identical with events or composed of them.

4.9 A Puzzle

The following not only seems reasonable but positively *de rigueur*:

> (TF[420]) The idea of there being free-floating property-instances makes no sense. Property-instances must belong to *things.* Surely there can't be free-floating instances of shape or mass or color. For there to be an instance of the property of being spherical, for example, there must be some *thing* that has that property. That property is either instantiated by some *thing*, e.g. a bowling ball, or it isn't instantiated at all.

But we've just seen that the following view, which is incompatible with TF, must be correct:

> (PB)[421] Things consist solely of property instances. "Substances" are either surds *or* they're causal relations holding among property-instances. If they're surds, they don't exist. If they're causal relations of just-described kind, then everything that exists is either a property-instances or is wholly composed of property-instances.

I don't know how to solve this problem, and I know of no one who has worked on it. But here's what I'd say if forced to hazard a guess, the operative word being "guess." Up to a point, TF is right. There can't be free-floating shapes or masses. Where there aren't things, there aren't shapes or masses. But, I would propose, the most *basic* property-instances, these being the ones out of which every occupant of space-time is constructed, are not instances of roundness or redness or of any of the other properties in terms of which we naturally understand the world. I propose the most fundamental building blocks of the constituents of space-time are instances of properties that, unlike instances of shape or mass, *can* coherently be thought of as free-floating.

5.0 Degrees of identity

It's generally held that, for any object *x* and any object *y*, *x* and *y* are either identical or they're not; they can't be identical to a certain degree.

Is this correct? Remember that we distinguished between diachronic and synchronic identity-claims. "Barack Obama is the current U.S. President" is a synchronic identity claim, and "I [JMK in 2009] am identical with that blobby infant [in the photo from 1973]" is a diachronic identity claim. Where synchronic identity claims are concerned, there are indeed no degrees of identity. *x* and *y* are synchronically identical or they aren't—there's nothing intermediate between total synchronic identity and total synchronic distinctness.

But diachronic identity is a different matter altogether. Diachronic identity not only can come in degrees, but *must* do so. Let "*x* = *y*" be an arbitrary diachronic identity claim. So far as it's true, it's only if the following conditions are met. First, *y* is a causal successor of *x*. In other words, there is some sequence of events, such that each event in that sequence is caused to occur by its predecessor (if it has one) and causes its successor to occur (if it has one), and such that *x* and *y* are both installments in that sequence. But for "*x* = *y*" to be true, it isn't enough that these conditions be met. If *x* is the particular gamete that, in due course, became me, and *y* is me, "*x* = *y*" isn't true.

But why not? Because *x* and *y* are *too* different. For "*x* = *y*" to be correct, it is necessary not only that *y* be *x*'s successor in some causal sequence, but also that the relationship borne by *x*'s condition with respect to *y*'s condition be a predictively and explanatorily robust one. I am structurally *so* different from the gamete that eventuated in me, and the bridges that had to be crossed in order for that gamete to eventuate in anything at all like myself were so rickety, that the relationship borne by that gamete to my present self isn't that of young JMK to old JMK, but is more like that of ancestor to descendant. To be sure, a single entity can change a great deal. The caterpillar is structurally very different from the butterfly it becomes. But the predictive and explanatory relations borne by the caterpillar with respect to the butterfly are sufficiently tight enough that it's more fruitful to see the change in question as internal to the existence of some one thing than it is to see it as a case of one thing's being replaced by some distinct thing.

Explanatoriness and predictiveness come in degrees. Since "*x* = *y*" is true

to the degree that *x* bears the relevant explanatory and predictive relations to *y*, thinghood is a matter of degree. It's also a matter of context. In some explanatory contexts, it *would* be correct to say of some gamete that it is identical, in the diachronic sense, with some philosophy professor. These points are developed in Sections 7.0–7.5 of the present chapter, which is concerned with the conditions that two different states of affairs must satisfy if they are to both be constitutive of some one entity.

5.0.1 Degrees of personal identity (continued)

We tend to think that *personal* identity either holds or it doesn't. "I remember *being* little Timmy," it will be said. "I remember playing kickball with my friends; I remember learning how to read; etc. If I remember *being* Timmy, then I *am* Timmy. It makes no sense to say that I might be 83% identical with so and so."

On the contrary, it makes plenty of sense. True—you have various memories (of playing kickball, watching cartoons, waiting for the ice-cream truck to come to your block, etc.) But all that shows is that certain currently existing mental structures and events are derivatives of the mental structures and events associated with some body (literally, some *body*) that is related to the body you now occupy in the way that a caterpillar's body is related to that of the butterfly into which it metamorphoses.

We're dealing with relations of succession and causation—nothing more. There isn't some mysterious inner flame, some underlying substance, that remains untouched throughout it all and that somehow underlies the fact that you and the little kickball player are identical. (Our tendency to believe otherwise is an expression of our need to condense vast, intricately structured constellations of events into homogenous units.[422]) Even if there is such an inner flame, it is inert *except* in so far as it facilitates the relevant psychological and physiological events. But so far as it does facilitate them, those events do all the work—they are what "keep you going," so to speak. The inner-flame is a free wheel and thus drops out of the picture. Thus, so far as you are identical with the little kickball player of yesteryear, it is because you and he are connected to each other by a sequence of events whose earlier members provide information about, and provide the causal basis for, its later members.

"If you're right," it will be asked, "why are we so convinced that diachronic identity is absolute—that I either *am,* or am not, the little kickball player?" We are not so convinced. What does a person even mean when he says that he *is* the little tyke in the photo? Does he *mean* that there is some underlying, unphased soul-substance that inhabits that little tyke and also inhabits him? It's hard to see how this could even *possibly* be what he means: since the terms "substance," and therefore "soul-substance," don't mean anything, so is the sentence that we are identical with "soul-substances"; so, in saying it, one cannot possibly mean anything (unless of course one is

speaking in an ironic or otherwise non-literal fashion). So, when a person says that he *is* the little tyke in the photo, he must mean that there is some sort of *connection* between himself (i.e., his current self) and that little fellow; and that connection, obviously, is one that consists, at least in part, in his being the outcome of a series of changes the first of which was undergone by that little tyke. (As we'll soon see, those changes must satisfy certain conditions; otherwise, instead of constituting the little tyke's developing into an adult, they'll constitute his ceasing to exist.)

What, more specifically, is the nature of this connection? (To sidestep unnecessary complications, we'll assume that the little tyke in the picture *is* the person who is alleging that he is that little tyke.) There is obviously a causal dimension to it. That person's current self is the end-*result* of the series of changes first undergone by the little kickball player.

But that connection cannot be of strictly causal nature: the relevant series of changes must form a coherent narrative. One's adult self inevitably has many beliefs, intentions, and values that his four year old self did not have. Nonetheless, setting aside the terminal links in the sequence of episodes composing one's life, each such link must have a certain continuity with its predecessors and its successors. What one is doing on Monday is the result of an intention one formed on Sunday; and, depending on whether one's beliefs as to how fruitful one's Monday-efforts were, what one is doing on Wednesday is either a continuation of those efforts, a modification of them, or an abandonment of the intention driving them.

Thus, what unifies various different psychological states into a single person is that they are *agentially unified.* Each state must represent the initiation, execution, modification, or disintegration of some single endeavor. Selves are the product, not the basis, of intentions---and therefore of the desires, beliefs, and values underlying those intentions. Language misleadingly suggests that it is the other way around. "First there are people," it suggests to us. "Initially these people are empty vessels. But they come to be occupied by beliefs, values, and the like. An intention must belong to some existing *person.* There can't be orphan intentions."

The thinking underlying this argument involves a subtle but hideous fallacy. This fallacy resembles, and is actually an instance of, the fallacy embodied in the idea that space-time is an empty vessel that exists independently of the events that occupy it.

5.0.2 An illustration of the principle being defended

A brief detour through metaphysics may help loosen the resistances that people have to the analysis of selfhood that I wish to defend. (This detour will end at the beginning of Section 5.0.6. I say this so as to enable readers who don't need to make this detour to skip it.) It makes no sense to suppose that, although *x*'s causal relation to *y* is qualitatively identical with *z*'s causal relation to *w*, *x*'s spatiotemporal relation to *z* differs from *z*'s spatiotemporal relation to *w*. It makes no sense to say that, although *x* and *y* are causally related to each other in exactly the way in which *z* and *w* are causally related to each other, *x* is 100 feet from *y*, whereas *z* is 10 feet from *w*. For the distance between *x* and *y* to be 10 times greater than the distance between *z* and *y is* for the one pair to be causally interrelated in a way in which the other pair is not causally interrelated; it *is* for a causal signal emitted by *x* to take 10 times longer to reach *y* than it does for a causal signal emitted by *z* to reach *w*; it *is*, other things being equal, for radiation sent from *x* to *y* to compare unfavorably in intensity with the radiation sent from *z* to *w*. To say of one event that it bears spatiotemporal relation *R* with respect to some other event *is* to say that the former bears a certain causal relation to the latter; it *is* to make a claim as to the conditions that *x* must fulfill if it is to have a given effect on *y*.

One can of course *imagine* two pairs of objects, one of which consists of objects that seem close to each other and the other of which consists of objects that seem remote from each other. And, with that image in one's mind, one can stipulate that the members of the one pair are causally interrelated in the same way as the members of the other pair. But this argument undermines itself. In imagining that the members of the one pair appear closer to each other than do the members of the other pair, so far as one's thinking is coherent, one is assuming that, other things being equal, the light-signals transmitted to one's eyes by the objects that appear to be close to each other differ from those transmitted to one's eyes by the objects that appear to be far from each other. And, so far as one's thinking is coherent, one is therefore assuming that, other things being equal, the light traveling by way of one's sensory surfaces between the objects that appear to be close to each other is different from the light traveling by way of one's sensory surfaces from the light traveling between the objects that are remote from

each other, in some way or other. Thus, one is assuming that, if one of the two objects that appear to be close to each other is to have a given effect on the other (viz. that of disturbing the latter by striking it with a light signal that first bounces off of one's eyes), what it must do to have that effect differs from what one of the other two objects must do to have a comparable effect on *its* mate.

It follows that the causal structure of the world is more basic than its spatiotemporal structure. If it assumed that *x*'s causal relations to *y* are identical with *w*'s causal relations to *z*, it makes no sense to say that *x*'s distance from *y* is different from *w*'s distance from *z*. We can change the causal relations without changing the distance between them (e.g., we can give *x* a positive electric charge without bringing it any closer to *y*); but we cannot change the distances between *x* and *y* without changing the causal relations between them.

This entails that spatial relations (and, for exactly similar reasons, temporal relations) are derivatives of causal relations. A brief discussion of the concept of *supervenience* may help. Two people who are microphysically identical cannot possibly be biologically different. (Two biologically identical people won't cease to be biologically identical if there is some isolated microphysical change in the one person's body that has no counterpart in the other person's body.) But two people who are biologically different must be microphysically different. (If the one person has a heart-attack and the other does not, all manner of microphysical changes occur in the one person's body that don't occur in the other.) Thus, the biological, though different from the microphysical, is "fixed" by it; to use the contemporary jargon, the biological "supervenes" on the microphysical. Given that the biological cannot vary without the microphysical varying, but that the microphysical can vary without the biological varying, it follows that there is some very important sense in which the microphysical is more basic than the biological.

Just as the biological cannot be changed without changing the microphysical, so spatiotemporal relations cannot be changed without changing causal relations. And just as the microphysical can be changed without changing the biological, so the causal relations can be changed without changing spatiotemporal relations. Thus, just as biological supervenes on the microphysical, so spatiotemporal relations supervene on causal relations.

5.0.3 An illustration of this principle (continued)

There is a clock resting on top of body *x*. Bodies *y* and *z* are far from *x* and also from each other. We will leave it open for now whether the distance between *x* and *z* coincides with either the distance between *x* and *y* or between *y* and *z*. Apart from containing *x*, *y*, and *z*, the region of space containing those three objects is a hard vacuum (i.e. it's perfectly empty). So there is nothing *other* than the distances involved to explain why a given light beam would take longer to travel from one of those objects to one of the other two. At time *t*, *x* simultaneously emits two light beams. The clock on top of *x* duly notes this fact. The one light beam heads towards *y*, and the other heads towards *z*. The clock's way of measuring time is to bounce a light-beam in between mirrors at opposite ends of its interior. (Every time one of the mirrors is struck, the structure of the mirror is altered just enough to leave a trace, so that it's possible to count the number of hits that occur during a given interval, but not enough to have a non-negligible effect on the reflective properties of the mirrors.) After striking *y* and *z*, respectively, the light beams emitted by *x* return to *x*. But those light beams don't return at the same time. During the interval between *t* and the time of the return of the *y*-bound light-beam, the light beam inside the clock has made ten billion round-trips. During the interval between *t* and the time of the return of the *z*-bound light-beam, the light beam inside the clock has made a hundred billion round-trips.

Obviously *x* is closer to *y* than it is to *z*. But the fact that the clock-internal light-beam made more round-trips during the one light beam's journey than it did during the other is not a *consequence* of the fact that *x* is closer to *y* than it is to *z*. It's *what it is* for *x* to be closer to *y* than it is to *z*.

5.0.4 Some qualifications

Of course, there are different ways of measuring distances. Light rays, and causal signals generally, needn't be used, and are hard to use, where small distances are concerned. One doesn't use light-waves, or even sound waves, to measure the length of a swimming pool. Instead, one uses a measuring tape, or one counts the number of times that a rod laid end-wise fits into a distance parallel to that of the side of the pool.

But this only reaffirms our point. Supposing that D_1 and D_2 are the end-points of the side of one pool, and that D_3 and D_4 are the end-points of some other, it is by performing operations that relate the *causal* properties of the one pool's end points to the other's that one determines the relative distances between those two pairs of end-points. In connecting D_1 and D_2 with, say, a measuring tape, one is *ipso facto* transmitting a causal signal from D_1 to D_2, and the same is true if one uses a meter-rod instead. The distance between D_1 and D_2 is greater than that between D_3 and D_4 if, other things being equal, the relevant meter-rod must be moved *more* times in the first case than in the second----or, supposing that measuring-tape is being used----if, other things being equal, the one measuring-tape exceeds the other in respect of how great the post-measurement distance is between its two end-points.

A given magnitude (e.g., a given distance) can typically be measured by means of different causal mechanisms. (A given distance can be measured using light-rays, meter-rods, measuring-tapes, etc.) Some might take this to show that distances are not causal relations. This is precisely the conclusion *not* to draw. The conclusion to draw is that distances are abstract causal relations. More precisely, they are commonalities holding among causal relations.

5.0.5 These points in relation to the measurement of time

Everything just said about the measurement of distance holds *mutatis mutandis* with respect to the measurement of time. For one instant in time to precede another by a greater amount than it precedes some third interval *is* for there to be causal differences between those intervals; it *is* for there to occur more periodic processes during the one interval than during the other. It is to be noted that a periodic process (light bouncing between two mirrors, an atom rotating, a measuring tape being stretched to connect two points in space, a yard-stick being moved from one place to another) *is* a causal process. So the idea of a space-time that is bereft of events is absurd, since space-time relations are higher-order causal relations; i.e. they are commonalities holding among causal relations holding among states of affairs.

5.0.6 The bearing of the points just made on the nature of diachronic personal identity

Exact analogues of these points hold with respect to personal identity. The idea of a vacant-mind---one devoid not only of specific mental occurrences, but also of the enduring psychological structures that give rise to such occurrences---is absurd.

Further, supposing *per impossibile* that such a nullity could exist, it would have no effect on one's psychobiography. Given two minds, each initially an empty vessel, any difference in the mental lives of the owners (so to speak) of those minds is a function solely of what happens in them---of what events occur in them and of what structures come to constitute them. This shows that what we call minds are much like spaces. We tend to think of a vessel that happens to be occupied but needn't be. But that's not the right view. An event-free space is a surd. To the extent that it even makes sense to say that such a thing as space exists, it must be identified with the totality of *spatial relations* holding among events. Put incoherently but graphically, what happens "in" space *constitutes* space. Put less incoherently, but no less graphically, space is no more independent of the events that we (misguidedly) describe as being "in" it than a table is independent of the molecules that compose it.

Space is distinct from events *only* in the sense that that the structure of space is determined by the *interrelations* holding among events, and not by the specific identities of those events. The same thing *mutatis mutandis* holds of tables. That is why, if *M* is one of the molecules composing a given table, replacing *M* with some other molecule that occupies the same relative position as *M* won't destroy the table. It is *only* in the way in which tables are distinct from the molecules that compose them that they are distinct from the thoughts, and thought-generating structures, that stock them.

5.1 The limits of Leibniz's Law

There's no doubt that Leibniz's Law correctly models the logic of *synchronic* uses of expressions such as "identical with." Given that Barack Obama is identical with the Commander-in-Chief, Leibniz's Law correctly warrants the inference that the Commander-in-Chief has a given property iff Barack Obama has that property.

The orthodox position is to say that Leibniz's law *also* governs the *diachronic* use of such expressions. But this is most doubtful. Right now I'm a professor. The blobby infant in the picture I'm now holding is not a professor. So I have a property that it does not have. Therefore, by Leibniz's Law, I am not it.

5.1.1 Time-indexing

Nonetheless, identity-theorists typically say that the blobby infant and I *do* have the same properties. To make this idea work, they bring in the highly artificial notion of a "time-indexed" property. I have the property of being-a-philosophy-professor-in-2009, and so does the blobby infant. The blobby infant has the property of being-preverbal-in-1973 and so do I.[423] Therefore, the fact that, unlike my six month old self, I am not *now* a blobby infant, being instead a non-blobby professor, does not make me different from my six month old self. Given any other property that I now have that I used to lack, or *vice versa*, a similar argument shows that I have *always* had that property. For this reason, it is said, Leibniz's Law *does* govern *diachronic* uses of expressions like "identity," "is" (in the non-predicative sense), and so on.

5.1.2 Time-indexing not a source of insight into questions concerning identity

In addition to being obviously contrived, this view has the absurd consequence that nothing changes. Whatever properties I have now, I've always had them and always will have them. In general, nothing acquires properties and, consequently, nothing ever loses a property. (Any case of property-acquisition is a case of property-loss. When a person acquires the property of being able to tie his shoes, he loses the property of being incapable of tying his shoes.) So if nothing ever acquires properties, nothing changes. Given that there is change, the no-change must be false.

5.1.3 Four-dimensionalism

To move forward, we must say a word about the debate that rages between those who accept a doctrine known as “four-dimensionalism” (henceforth, “4-D”) and those who reject that doctrine. According to 4-D, any given object is identical with a “space-time worm.” A space-time worm is an entity that is spread out in both space and time. Thus, according to 4-D, you are identical with a space-time worm that comprises all of the events and states of affairs constituting you from the moment of your birth (or conception) till the moment of your demise.

Bearing this in mind, let *t* be any instant in time during your lifetime. If 4-D is right, some *segment* of you exists at *t*; but *you* do not exist at *t*, since you, according to 4-D, are spread out over a period of (say) 80 years. So if 4-D is right, I cannot see you or shake hands with you; I can see and shake hands only some segment of you. The thing I shake hands with, says the advocate of 4-D, is some *time-slice* of you.[424]

In my view, I *can* shake hands with people, and also see them, play tennis with them, and otherwise interact with them. I thus reject 4-D.

5.1.3.1 Four-dimensionalism (continued)

Things change. 4-D denies this. If 4-D is right, the thing that is reading this sentence right now isn't *you;* it's just some time-slice of you. And the slightly different thing that was reading the first sentence of this chapter isn't you either; it's just another time-slice of you. The same is true of the thing that decided to buy this book a few weeks ago. According to 4-D, although the thing reading this sentence is different in some ways from the thing that was reading the first sentence of this chapter, those differences aren't expressions of the fact that some one thing has *changed.* According to 4-D, *you* don't change. For a given object to change is for the way that object is at one time to be different from the way it is at some other time. According to 4-D, for any time *t*, what exists at *t* isn't an object, and is only some "time-slice" of an object. Therefore, there are no two moments of time *t* and *t** such that how some one object is at *t* is different from how that same object is at *t**. Therefore, if 4-D is right, *things don't change.*

But this is absurd. Things change. I have changed, and so have you. And each of us will continue to change. 4-D is either dead wrong or it amounts to nothing more than a proposal that we use the words "thing" and "change" in non-standard ways. And even if by some remote chance it turns out that things *don't* change, it would be bad methodology to make our theory of personal identity depend on that turning out to be the case. In doing so, we'd be assuming facts (or non-facts, rather) that aren't yet in evidence.

Assuming, as we probably must, that things change, the only way to go is to say that identity, in the diachronic sense, is to be understood in terms of relations of causal connectedness—that statements like "Bill is the guy in the photo who is wearing a rugby shirt" are to be analyzed in the way described earlier. Since the relevant sort of connectedness obviously comes in degrees, so does identity.

It should be pointed out that David Hume (1711–1776), whose views we'll discuss in Section 6.0, takes a similar view, and so does Derek Parfit (1942-). [425]

5.1.4 One possible source of the confusion underlying acceptance of four-dimensionalism

It should also be pointed out that those who accept 4-D very much seem to be reading metaphysics off of sentential surface structure. The expressions we use to denote the relation of diachronic identity are homonyms of the expressions we use to denote the relation of synchronic identity. Given *only* that fact, along with some obscure and ambiguous intuitions about selfhood, 4-dimensionalists infer that synchronic identity *is* diachronic identity. That inference is completely spurious; and it's one that no self-respecting analytic philosopher would make, given that the essence of analytic philosophy is a stubborn refusal to take surface structure to be a foolproof guide to truth.

5.2 An identity-related paradox and the need for a diachronic conception of identity

A story will make it clear that that 4-D is simply too revisionist to take seriously. At time *t*, person X_1 splits, like an amoeba, into two people, X_2 and X_3. X_2 is just like X_1—in any case, X_2 resembles X_1 as much as you, in your current condition, resemble the person you were a minute ago. And X_3 resembles X_1 to the same degree. Further, the causal connection that X_1 bears to X_2 is qualitatively identical with the causal connection that X_1 bears to X_3.

It's obvious that $X_2 \neq X_3$. After all, they're in different places. And it would obviously be arbitrary to say that X_2, but not X_3, was X_1. Whatever claim X_2 can make to being X_1, X_3 can make precisely the same claim. David Lewis' solution is to say that X_2 and X_3 have been there all along: as soon as X_1 came into existence, so did X_2 and X_3. There were three people rolled into one body and one mind.

Lewis' position is absurd. Before he split into two, X_1 was no more a compendium of different people than you are. There is no relevant psychological or physiological difference between X_1 before the split, and anyone else. If X_1 had died before the split, not even a shadow of a case could be made that "he" was really a "they"—that he was really three people. So, supposing that X_1 *does* live and, in due course, splits into X_2 and X_3, his undergoing that split *retroactively* puts two extra people in X_1's body. So Lewis' solution goes through only if there is retroactive causation.

That is bad enough, but it gets worse. Even though, if Lewis' solution is to go through, there must be retroactive causation, that causation must fail to make any psychological or physiological difference to X_1 or, for that matter, to the state of the universe before X_1's split. But to be a cause *is* to create change. So Lewis's position collapses into the absurdity that there are causes that aren't causes.

5.2.1 An identity-related paradox and the need for a diachronic conception of identity (continued)

It's obvious that Lewis is jumping through hoops. He's working with a broken concept and is trying to cover up the cracks in it by mutilating wholesome and intact concepts. But Lewis' position is *de rigueur* if one holds that the sense in which adult-JMK is "identical" with baby-JMK is identical with the sense in which 2 is identical with 1 + 1. We've already seen why. X_2 is no less "identical with" X_1 than is X_3. And X_2 is no less identical with X_1 than you are identical with the person you were a minute ago. Therefore, *if* we say that you are "identical" with that person in the same sense in which 2 is identical with 1 + 1, we must say that *both* X_2 and X_3 are identical with X_1 in the very same sense. Since it's a datum that X_2 and X_3, being in different places, are distinct from each other, it follows that, just as Lewis says, the mind and body had by X_1, prior to the split, were shared by X_2 and X_3. It follows, in other words, that X_1, X_2, and X_3 were there all along, all rolled into one package.

Given how preposterous that is, it's clear what we must say. X_2 is "identical" with X_1 in the sense that it is X_1's successor; the same is true of X_3. And the same thing *mutatis mutandis* is therefore true of you. You are identical with your teen self in the sense that you are that entity's successor.

True—the expressions we use ("identity," "identical with," etc.) to describe the relationship between Barack Obama and the current President, or between 2 and 1 + 1 are the same ones we use to describe your relationship to your teen self. But so what? Language is mirror of thought; thought isn't always perfect. Thought is especially prone to error when it comes to emotion-laden issues, such as personal identity. Human beings don't want to believe that they are event-sequences that can become frayed, like an old rope, and dwindle into nothingness. Human beings don't want to believe that their identities are internal to emotion-driven narratives. But what we want to believe is irrelevant. What *is* relevant is that, if we take the conventional view, then we must follow Lewis in embracing the absurdity that you could be 80 different people, even though you are physiologically and psychologically just like somebody who is but one person.

6.0 Personal identity revisited: Hume's analysis

Few have done as much to clarify the structure of the psyche and nature of personal identity as David Hume (1711–1776). Hume had a mentalistic conception of personal identity. In his view, there are no people where there are no minds. But, unlike us, he held that there are no minds where there are no occurrent thoughts—where, in other words, there is no mental *activity*.

Why does Hume hold this? Because Hume is a mentalist *who doesn't believe in the existence of structures of any kind.* Hume acknowledges that there are events; but he denies that there is anything underneath those events that guide them. There are no forces, no structures, no powers. (See Chapter 17.) And this is why, being a mentalist, he must identify the self with a stream of mental events.

There is an obvious objection to this view: "If Hume is right, then the person who wakes up in the morning isn't identical with the person who went to sleep at night." Hume is aware of this objection, and his response to it is very profound. He says that diachronic identity is pseudo-identity. The concept expressed by the occurrence of "is identical with" in "JMK is identical with the blobby infant in the photo" simply isn't the same concept as the one expressed by its counterpart in "Barack Obama is identical with the current U.S. President." Taking it for granted that synchronic identity is a genuine form of identity, he infers, very reasonably, that diachronic identity is pseudo-identity.

6.1 A preliminary point: Hume's empiricism

Empiricism is the doctrine that everything that is known is known through sense-perception.[426] So far as empiricists grant that anything can be known that isn't directly observed, they hold that it is inferred from what is directly observed *provided* that the inference-rules involved are themselves known from direct observation alone.

No sooner is empiricism defined than its coherence is called into question. An inference-rule is a *normative* truth. It tells you what you *ought* to infer from a given premise. But so far as it apprises you of anything, observation apprises you of what *is*, not of what *ought* to be.

Unlike other self-described empiricists, Hume didn't help himself to any rules of inference whose legitimacy doesn't have a strictly perceptual basis. This was because, unlike his empiricist brethren, he had the acuity to see that many inference-rules that seem to be empiricism-friendly ones are not. Hume's empiricism was doubly strict. First, he had extremely conservative views as to which rules of inference had the requisite perceptual basis. Second, he had extremely conservative views about what sense-perception *directly* tells us.[427] What Hume says about personal identity must be understood in terms of the fact that he is a very strict empiricist.

6.1.2 Hume’s first argument

Hume gives two arguments for his contention that selves are composed of the mental contents that occupy them. Here is the first:

> For my part, when I enter most intimately into what I call *myself,* I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I can never catch *myself* at any time without a perception, and never can observe anything but the perception. [428]

There are two ways to interpret this. On the one hand, Hume could be saying that selves don’t exist. On the other, he could be saying that selves are *constituted* by the thoughts, feelings, etc., that we describe them as “having.”

The right interpretation is the second one. Notice how frequently the word “I” occurs in his argument. He obviously believes that, when he uses that word, he's referring to himself and thus to his self. And surely he's right about this. It seems a truism that, if they refer to anything, tokens of "I" refer to selves; and it also seems a truism that tokens of "I" *do* refer. Also, in other parts of the Treatise, Hume talks at length about the relation that two different mental contents must bear to each other to belong to the same mind or self. Thus, Hume’s position is that, although selves do exist, they are no more distinct from the mental events that occupy them than a swarm of wasps exists independently of the wasps that compose it. For an event to *occupy* a mind is for it to *constitute* it.

The just-quoted argument, it will be noticed, is *empirical*. Hume is arguing on observational, as opposed to strictly logical or conceptual, grounds that there isn’t anything that underlies his various thoughts. When he looks inward, he finds thoughts and nothing else; in particular, he doesn’t find a non-thought that “supports” these various thoughts. This fact is important, as we’ll now see.

6.1.3 Problems with Hume's argument

Given only that we can't observe something, it doesn't follow that it doesn't exist. There are many unobservable entities (e.g., electrons). If Freud (1915) is right, there are unobservable (unintrospectible) mental entities (e.g., unconscious urges). Saying that selves don't exist since we can't introspect them is like saying that electrons don't exist since we can't see them.

A related possibility that Hume is ignoring is that minds are *structures* of the kind described earlier. (This oversight is probably deliberate, as we'll see in a moment.) He's also ignoring the correlative possibility that specific mental occurrences (e.g., images, tickles) are *products* of minds, and are not themselves *constitutive* of them.

In ignoring these possibilities, Hume is stacking the deck in his favor. We are pretheoretically inclined to see minds as being to specific tickles, itches etc., what car factories are to cars. In other words, we're inclined to see minds as being the things that give rise to specific mental contents, and we're therefore inclined to see minds as *not* being constituted by those contents. (In this context, by "specific mental content, I mean *occurrent* mental content---e.g., a mental event, *as opposed to an enduring mental structure.*)

In any case, we're pretheoretically much more inclined to see minds in this way than we are to see them in Hume's way, viz., as sequences of mental occurrences, none of which has any moorings in any structure and each of which, therefore, just happens. And in simply dismissing that possibility out of hand, Hume is question-beggingly assuming the falsehood of what is probably the most plausible rival to his own view.

In any case, supposing, as Hume does, that such structures aren't known through introspection, it isn't necessary to conclude that they don't exist; for it would be equally reasonable, if not considerably more reasonable, to conclude that introspection is a limited medium.

6.1.4 Why does Hume deny that there are mental structures?

Hume is a hardcore empiricist. You know it only if you (i) observe it or (ii) observe direct evidence for it. (*x* is "direct" evidence of *y* if, given no inference rules whose legitimacy cannot be established strictly on the basis of observation, *y*'s existence can be inferred from *x*'s.) Hume believes that mental structures cannot be known in either of these two ways. (He's right.) So they don't exist as far as he's concerned.

Hume doesn't believe in structures of *any* kind. A structure is something that underlies events and guides them. Hume doesn't believe that anything guides anything. He doesn't believe that anything *makes* anything happen or otherwise *influences* anything. In his view, words like "force," "power," "coercion," "guide," and "regulate" are meaningless. He holds this because, so he believes, forces cannot be observed and neither can direct evidence of them. (Hume's views on causality are the subject of Chapters 15 and 16.)

Thus, Hume's empiricism blocks him from granting the existence of mental structures, given that such structures can't be known (para)perceptually. Second, given Hume's very reasonable view that one can't observe forces or any direct evidence of them, his empiricism prohibits him from granting the existence of structures of *any* kind. Thus, Hume's empiricism makes it doubly impossible for him to grant that there is anything to one's mind other than fleeting, conscious events. (In Chapter 15, we saw that, contrary to what Hume holds, forces can be directly observed---that, in fact, *everything* one observes is a force and, moreover, is observed *as* a force.)

6.1.5 Problems with Hume's argument (continued)

In an attempt to show that there isn't anything to one's mind other than the various specific mental occurrences that it houses, Hume says that, when he looks inward, he doesn't find an inner flame; all he finds, he says, are specific images, tickles, etc.

But unless this inner flame is *itself* a fleeting mental occurrence—unless it's the sort of thing that can be disclosed to one in the form of an introspectible image or sensation—nothing significant follows from the fact that, when he introspects, he doesn't find this inner flame. So for Hume's argument to go through, it must be assumed that the self, if existent, is to be known to one in the same way as tickles and itches. It must be assumed, in other words, that the self is a fleeting conscious occurrence. But that's the very thing Hume is trying to show; and, were Hume to assume it, his argument would be viciously circular.

6.1.6 The problems with Hume's argument are reflections of his hardline empiricism

It is through images that our perceptual modalities (sight, touch, etc.) tell us about the world. (Not all images are visual; there are acoustical images, tactile images, etc.) Your seeing a tree consists in your having an image of that tree; your feeling the tree with your hand consists in your having a tactile image of that tree; and so on. There's no perception where there's no image. It follows that, if introspection is an inwardly directed perceptual modality, it tells one nothing when it isn't giving one an image.

Hume says that, when he looks inward, he encounters various "perceptions" (images), but no self. On this basis, he says that there is no self. But the right conclusion is that, *if* there is a self, it isn't disclosed to one in the form of an image.

Supposing that it exists, there are, ultimately, two possibilities as to what the self is:

(i) it could be one's center of awareness;

or

(ii) it could be some structure underlying the events in one's mind and guiding them.

(i) and (ii) are not mutually exclusive. So a self could be a person's center of awareness that is also a structure underlying and regulating that person's mental activity.

Given that either (i) or (ii) describes what the self, supposing it to exist, must be, it follows that the self cannot be known in a way that can be modeled on sense-perception, and there is thus no merit to Hume's paraperceptual demonstration of the non-existence of the self.

It is uncontroversial that we are directly, non-inferentially aware of many mental entities that we *don't* know of in some image-mediated, perception-like way. I know I believe that 1 + 1 = 2. It's obviously not through *inference* that I know this—I know it directly. And it obviously isn't through sense-perception that I know it. Finally, it isn't by virtue of my experiencing, or being aware of, some *mental image* that I know it. *But surely I know it.* So I

know it directly. It's a pretheoretic datum—not the result of an intellectual working over of raw data. No—it *is* one of the raw data.

The same is true of many other mental contents of mine. I know that I intend to write about causation later today. It obviously isn't through *inference* that I know this—it isn't because I psychoanalyzed myself or engaged in any other such artifices. And it obviously isn't through sense perception (sight, hearing, etc.) that I know it. Nor is it through the inwardly directed paraperceptual modality that Hume describes. But I know it. I know it in the same way that I know that 1 + 1 = 2, that I regret not calling my mother on her birthday, that I plan on improving my skills as a pianist, etc.

I know these things through some faculty that gives me direct, non-inferential knowledge of at least some facts about my own mind. Were Hume to deny the existence of this faculty, he'd be butchering obvious facts to make his argument go through and his position would be a most unempirical one. It is thus an empirical fact that not every fact about one's mind that one is non-inferentially aware of is disclosed to one in the form of an image; and it's thus an empirical fact that not all knowledge is arrived at through perception or through anything at all like it. It is thus an empirical fact that empiricism is wrong!

6.2 Hume’s second argument: a preliminary point

Hume’s first argument is presented as an empirical argument. He says that his observations yield no evidence that there is some central fire around which his various thoughts orbit. His wording, and the wording that it is natural to use when paraphrasing his argument, suggests that Hume is trying to establish the non-existence of the self on grounds comparable to those on which an astronomer would try to establish the non-existence of a certain planet. “If I had evidence, I’d believe. But I don’t, so I don’t.”

But so far as self-observation yields *any* relevant evidence that evidence seems to suggest that *there does exist* within oneself a center of awareness. When I self-observe, I *do* feel that there is a core-being around which everything else orbits.

This feeling of mine is by no means probative. I don’t know that it’s accurate. I don't know that it’s a *correct* observation. In fact, I don’t know that it’s an *observation* at all. It may be that, in having this feeling, I’m experiencing as an empirical datum what is in fact a projection into the observational arena of my acceptance of an erroneous, theory-driven *interpretation* of the relevant data. In fact, this is Hume's position. Hume explicitly says that those who, like myself, claim to be aware of an inner flame are either lying or they’re in the grips of some preexisting, empirically unfounded theory of what the self *must* be.

The problem is that, even if, as Hume claims, my feeling that there’s an inner flame is an illusion powered by a very wrong theory, it’s a datum that this *feeling* exists. And it *isn’t* a datum, though it may be a truth, that this feeling is an illusion. So, even though Hume’s opponent may be wrong, it’s *Hume,* not his opponent, who’s making the theoretical point. This suggests that he has some theory in mind, his acceptance of which, and not some body of observations, led him to accept the “no inner flame” view.

6.2.1 Hume’s second argument (verbatim)

And, in fact, Hume does have just such a theory (I quote):

> When my perceptions are remov’d for any time, as by sound sleep; so long am I insensible of *myself*, and may truly be said not to exist. And were all my perceptions remov’d by death, and cou’d I neither think, nor feel, nor see, nor love, nor hate after the dissolution of my body, I should be entirely annihilated, nor do I conceive what is farther requisite to make me a perfect non-entity.[429]

This argument, unlike the first, isn’t empirical at all. This time, Hume is trying to show on purely conceptual grounds that there is nothing underlying the various events occupying a person’s mind. He’s arguing that the very *idea* of such a thing is an incoherent one, like that of a square circle. Why is this idea incoherent? Because, says Hume, it has the absurd consequence that there could be thought-free minds.

6.2.2 Evaluating this argument: Hume's failure to distinguish between two very different meanings of the word "thought"

Hume is right that *if* there is some inner flame *underlying* one's thoughts, and therefore distinct from them, it *does* follow that one's self could exist in a thought-free condition. But Hume is wrong to say that a thought-free self is an absurdity. The reason Hume *thinks* it absurd is that he fails to distinguish two very different meanings of the word "thought."

Sometimes the word "thought" refers to mental occurrences—to what are now known as *occurrent* mental phenomena. An occurrent mental phenomenon is anything mental that happens. Sense-perceptions are occurrent mental phenomena; and so are acts of ratiocination and emotional epiphanies. (This list is far from complete.)

But sometimes the word "thought" refers to the relatively stable *structures* that *underlie* mental occurrences. "For the first twenty years of my academic career," a philosopher once said to me, "I thought that Platonism was false." This person was describing a stable, structural constituent of his mind, not a mental *happening.*

If, by "thought" Hume means "mental happening" then, contrary to what he alleges, there's nothing incoherent in the idea of a mind that, during some period, isn't populated by thoughts.

If by "thought" Hume means "stable mental structure" then Hume *is* right to say that there couldn't be thought-free minds. Earlier, we saw why this is so. If selves were distinct from their thoughts, there could be thought-free minds. So, given that there cannot be thought-free minds, selves are constituted by the thoughts that belong to them.

So, after a fashion, Hume is right. Just as he says, there is nothing to one's self other than the thoughts one has. But this claim of his isn't right in the way he thought it was. He thought that one's self consists of the thoughts (i.e., fleeting mental occurrences) that occupy it, and not of the thoughts (i.e., stable structures that create and regulate such occurrences). We've already discussed why Hume didn't even grant the existence of stable mental structures or, indeed, of structures of any kind.

7.0 Objectual identity revisited: Which causal sequences are thing-constitutive?

Earlier in this chapter, we saw that every object—every rock, tree, and vase—is a causal sequence. But not every causal sequence constitutes some one object. Some constitute more than one. Others constitute none.

Here's an example of one that constitutes no objects. Suppose that billiard ball #1 strikes hitherto stationary billiard ball #2, causing it to roll. Ball #2 then strikes hitherto stationary billiard ball #3, causing it to roll. And so on. This sequence of events doesn't constitute a single entity; this sequence isn't like a rock or a tree or an ice cube.

Here's an example of a causal sequence that constitutes more than one object. Let S_1 be the series of events constituting some seed *and* the process by which that seed turns into a tree and *the* tree itself. S_1 is a causal series—a textbook case. But it constitutes two distinct objects, not one. The seed isn't a mini-tree. It's a tree-precursor, not a tree, just as a gamete is a person-precursor, not an actual person. (Unlike people, gametes don't have thoughts or fall in love or play the harmonica.)

Two questions arise. What is the difference between those causal sequences that are thing-constitutive and those that are not? Second, what is the difference between thing-internal change, on the one hand, and wholesale transformation, on the other?

7.1 Some general features of thing-constitutive causal sequences

Nothing that endures for more than an instant is *not* a causal process. But what is distinctive about those causal processes that we think of as objects (e.g., frozen ice cubes) is that the changes constitutive of them occur within certain narrow and fixed limits.

There's nothing unchanging about any ice cube—even one that for the last fifty years has been inside in an undisturbed deep freezer. Given such an ice-cube, every particle constituting it is in constant flux; and the ice-cube itself is nothing but a flurry of activity.

But even though those particles are all moving vigorously, they're moving only in ways that allow the flurry they compose to retain more or less the same properties. The *flurry* doesn't change, even though the things composing it do. The reason the flurry doesn't change is that, although its components are changing incessantly, the way in which they change doesn't change.

Thus, what we *usually* describe as "changes" are really second-order changes: changes in the changes themselves. Each molecule composing a given ice-cube is an arena within which ceaseless change occurs. But so long as the ice-cube remains intact, there are no significant changes in the relations borne to one another by these different arenas of change. That's why the ice-cube is unphased by such changes. But if the ice-cube is smashed or melts into a puddle, those mutual relations *are* significantly changed. That's why the ice-cube does not survive such changes.

7.2 Change vs. obliteration

Not all objects are as static as ice cubes, and our analysis must be refined if it is to deal with more dynamic entities (e.g., a living human body). Like a rock or an ice cube, a living body consists of individual particles each of which bristles with activity. But where living bodies are concerned, there are changes, sometimes dramatic ones, in the relations that these arenas of change bear to one another. This happens whenever an organism develops in a positive direction, as when maturation or recovery strengthen it, or in a negative direction, as when old age or illness enfeeble it.

Thus, not all second-order changes jeopardize a thing's existence, and sometimes they're actually necessary for it. When do second-order changes destroy existence and when do they sustain it? "Such changes are existence-unfriendly," it might be suggested, "when they're abrupt or discontinuous and they're existence-friendly when they're not." But this isn't the right answer, since the melting of an ice cube, though decidedly thing-*un*friendly, is as smooth a process as the development of an organism, which is a decidedly thing-friendly process.

"Changes are thing-preserving," it will be said, "when they're *structure* preserving. The general morphology of an adult human being is similar to that of an infant. There are differences, of course, but they're internal to more fundamental similarities. Thus, the changes that connect the infant to the adult preserve structures. In general, thing-sustaining changes are structure-sustaining changes."

This position obviously has merit, but it won't quite do. There are cases where, without going out of existence, some one thing changes in fundamental structural respects. The butterfly doesn't resemble the caterpillar; but there is some *one* organism that is first a caterpillar and then a butterfly. So why is it that caterpillars that turn into butterflies continue to exist whereas ice-cubes that turn into puddles do not?

7.3 Dynamic integrity the essence of thinghood

Puddles by no means defy prediction. If you see a puddle that is expanding, you may know roughly how long it will continue to expand and roughly what its shape will be when it stops expanding. But in addition to being of limited scope and low quality, such predictions presuppose *ad hoc*, circumstance-specific knowledge. The predictive payoff is minimal, given the size of the epistemic investment.

By learning more about the relevant variables (e.g., the shape of the floor, the rate at which the puddle is expanding, etc.), you can increase the predictive payoff. But given a vast improvement in one's knowledge of these variables, there would be only a slight improvement in the predictive payoff. To have any chance of predicting with any precision what the puddle's shape will be when it stops expanding, one would need extraordinarily high-resolution knowledge. Further, interference-effects would guarantee that, no matter how good your measuring instruments were or how adept you were at using them, your knowledge of initial conditions would not exceed a rather low threshold. And supposing that, within these limits, you acquired as much knowledge as possible of the relevant variables, any predictions you could make on the basis of that knowledge would be of minimal scope. Given high-resolution knowledge of the puddle's state at a given time, you could make low-probability, low-resolution predictions about the puddle's state at times soon thereafter.

But the story is very different where butterflies, human beings, and other *bona fide*things are involved. Given a knowledge of slight disturbances to a caterpillar's anatomy, one can make high-resolution, high-probability predictions about the anatomical condition of the butterfly into which, in due course, it will metamorphose. Easily acquired low-resolution knowledge of an animal's state at a given time may yield a high-resolution knowledge of its state at a much later time.

Given a pregnant mother who consumes large enough amounts of alcohol, it can be predicted that, 50 years hence, the person to whom she gives birth (supposing that person to be alive) will be severely damaged, and it can be said with some precision how that person will be damaged. The other, brighter side of the coin is that, given easily acquired, low-resolution knowledge to the effect that a certain person, while in his formative years, is given certain kinds of stimulation, allowed certain kinds of freedom, etc., one

can make high-probability, high-resolution predictions about that person's condition in late adulthood.

Of course, butterflies and children aren't just things—they're organisms. And not everything that is true of organisms is true of all things. But points relevantly similar to the ones just made hold of non-organisms. A sculptor who is chipping away at a hunk of marble knows that, barring some extraordinary event, the shape had by the hunk when he's done chipping away at it is the shape it will keep for centuries to come. The people who built the Washington Monument knew that, barring some extraordinary event, it would remain standing, its shape and general physical condition more or less unchanged, for hundreds if not thousands of years to come. In general, a limited knowledge of a thing's state at a given time yields high-probability, high-resolution knowledge of that thing's state at a much later time.

Knowledge of this principle underlies all physical labor. It underlies the manufacturing of cars, houses, fishing rods, and elevators. All of this activity would be pointless were it not for the prediction-friendliness inherent in thinghood.

These predictions are defeasible, of course.[430] Well-built houses are destroyed by earthquakes. Well-educated children from loving homes turn out to be moronic ghouls. But it's *despite*, not *by virtue of*, its being a thing that a sculpture or child or anything else has a counterpredictive future. Embittered by his lack of artistic ability, some unknown kook breaks into the Louvre and destroys famous sculpture *X*. The world is shocked. Nobody could have predicted that the world would ever be without this masterpiece. *X*'s future was indeed an unpredictable one. But until its thing-status was completely undermined, *X*'s future *was* a predictable one. Thus, *X*'s future was unpredictable *only* to the extent that *X*'s thing-status wasn't determinative of it. So given only that *X*'s future was not predictable, it doesn't follow that things *qua* things aren't prediction-enablers; and given that *X*'s future *was* predictable to the extent that its thing-status had a hand in it, it follows that things *qua* things *are* prediction-enablers.

In general, to be a thing *is* to be a prediction enabler. It would, of course, be viciously circular to define a "thing" as a prediction- and explanation-enabling *thing.* But when we replace the second occurrence of the word "thing" in the last sentence with "complex of events," the circularity goes away. Thus, to be a thing is to be a prediction-enabling complex of events (or

“event-complex,” as we’ll say). By the same token, an event-complex that is predictively inert is, to that extent, a non-thing.

Prediction and explanation are conjugates of each other. An explanation uses known laws to go from effect to cause. (I know that Smith is now belittling Brown. I know that insecure, touchy people become angry when belittled. I know that Brown is such a person. So I predict that Brown will become angry.) The corresponding prediction uses the same laws to go from cause to effect. Thus, a swarm of events is a thing to the extent that it is prediction-and explanation-conducive.

7.4 Thinghood is a matter of degree

The property of being prediction-enabling comes in degrees, as does the property of being explanation-enabling. A brick house is richer in these properties than a plywood house; a plywood house is richer in them than a cloud; a cloud is richer in them than a breeze.

A consequence of our analysis is that thinghood comes in degrees. Some things are "thingier" than others. I accept this consequence. A brick house is more of a thing than a cloud; it's more "thingy." This means only that it provides a broader and more stable basis than does a cloud for prediction and explanation. Since this is what we are inclined to think anyway, it isn't to the discredit of our analysis that it has this consequence.

We are inclined to think that things must literally be solid, like rocks and ice cubes, and unlike puddles. So far as there is a connection between actual solidity and thinghood, it is that, within the narrow horizons within which we live, solid things typically provide a firmer basis for prediction than non-solid things. But non-solids can be as thingy as rocks. Though they're gaseous, as opposed to solid, stars are things to no less a degree than rocks. The reason is that, even though they're not solid, they have the dynamic integrity of things that, in terrestrial contexts, are typically had by solid things.

7.5 The contextual nature of thinghood

Despite its internal lack of cohesiveness, a galaxy can be highly thing-like in the context of an explanation of why certain extremely large-scale astronomical events are occurring—of why, for example, there is a marked tendency for objects in a certain very large region of space to head in a certain direction. And despite its *high* degree of internal cohesiveness, a rock can be as unthing-like as, in most contexts, gusts of air would be. To creatures whose minds moved millions of times more slowly than ours, rocks would be as unthing-like as gusts of air are to us. And to creatures whose minds moved millions of times more quickly than ours, gusts of air would be as durable, cohesive, and predictable as marble statues are to us.

Thus, a given system of events can be a thing—in certain respects, at certain junctures—in some explanatory contexts, while being unthing-like—in different respects, at those same junctures—in others. But that doesn't mean that things are created by our explanations.

8.0 Time-travel

It is debated whether time-travel is possible. Thanks to our analysis of what it is to persist in time, we can shed some light on this problem.

Setting aside the trivial and irrelevant fact that we travel forward in time at a rate of one minute/minute, the statement "time-travel is possible" has three distinct meanings:

> (a) One can "leap" from one time to some non-adjacent time; i.e., one can move in a *discontinuous* manner from one time to another.
>
> (b) One can travel forward in time at a rate of *more* than one minute/minute.
>
> (c) One can travel *backwards* in time.

It is generally thought that (a)–(c) are empirical questions that fall within the jurisdiction of physics. In actuality, they are questions of logic, not of empirical fact. Moreover, each of (a)–(c) is logically false and therefore couldn't possibly be correct.

Let's discuss (a) first. In this chapter, we have seen that anything that persists for any amount of time is a ***causal series.*** A causal series is a continuous series of states of affairs such that, given any two of them, one is the (in)direct cause of the other.

An illustration may be helpful. At any given moment during its existence, your car consists of an incessant flurry of mass-energy displacements; and your car is thus a series of "event-flurries," as we'll call them. Twenty-four hours ago, your car was constituted by one event-flurry. Right now it's constituted by a different event-flurry. (Let S_1 be the first event-flurry and let S_2 be the second.) And, of course, my car is currently constituted by some event-flurry. (Let S_3 be that flurry.)

Bearing all of this in mind, imagine the following. While your car was parked in your driveway, somebody crashed into it, severely damaging it. Your car therefore bears little resemblance to what it was twenty-four hours ago. It follows that S_1 and S_2 are qualitatively very different from each other. By contrast, S_1 and S_3 are as similar to each other as their being in different places permits. This is sheer coincidence: S_1 and S_3 are, we will suppose, as

causally disconnected from each other as any two non-simultaneous events on this planet can be. It follows that my car, as it is right now, is just like your car, as it was twenty-four hours ago, before it was damaged. It also follows that S_1 is much less similar to S_2 than it is to S_3.

Be all of this as it may, the car in your driveway is, whereas the car in my driveway is not, numerically identical with the car you owned twenty-four hours ago. Mediating between S_1 and S_2 is a continuous series of events such that, given any two of them, one is the (in)direct cause of the other. No such series mediates between S_1 and S_3. That's why (i) the car in my driveway isn't the very same car as the one you parked in your driveway twenty-four hours ago, their uncanny mutual resemblance notwithstanding. And that's also why (ii) the car sitting in your driveway is the very same car as the one you parked in your driveway yesterday, their *lack* of mutual resemblance notwithstanding. Obvious generalizations of these points establish our thesis that objects are identical with causal series.

A consequence of the fact that objects are such series is that a given object ceases to exist the moment the corresponding causal series is severed. If, instead of just being damaged last night, your car was vaporized, my car, though an exact duplicate of your (pre-vaporized) car, still isn't numerically identical with it.

Given that an object ceases to exist the moment the corresponding causal series is severed, it follows that, so far as it isn't logically false, the statement:

> (i) Smith disappears in 2009 and re-appears in the year 2023.

collapses into the statement:

> (ii) Smith is annihilated in 2009 and a Smith-*like* entity that isn't actually Smith *ex nihilo* comes into existence in the year 2023.

A question will help clarify the reasoning behind this last assertion: Where is Smith during that fourteen year interval? He's either nowhere or he's somewhere. Let's suppose that he's nowhere. In that case, for the reasons illustrated in connection with your car, whatever it is that pops into existence in the year 2023 isn't numerically identical with Smith. So let's suppose that he's somewhere—

that, at any given instant during that time period, he is in some place or other. But in that case he's no more time-traveling than anything else that exists during that interval. If he's next to some one tree during that entire period, he's no more time-traveling than that tree. If he's going from place to place, he's no more time-traveling than any one of the satellites that will orbit the earth during that period.[431]

So depending on whether he's somewhere or nowhere, Smith is either permanently annihilated in 2009 or he's no more time-traveling than anything else that exists during that period. Therefore, (i) is false.

To see why (ii) is false we need do little more than repeat some of the points just made. Suppose that Smith travels from 2009 to 2023 at a rate of one million minutes/minute. We've just seen that Smith cannot make this journey in a discontinuous manner; he cannot disappear from one time and reappear at some other, non-adjacent time. So Smith must make this journey in a continuous manner. So at any given moment during this interval, Smith must be in some place or other. It follows, as we saw a moment ago, that he is no more time-traveling than anything else that exists during that interval. The supposition that Smith is traveling forward in time at a rate of one million minutes/minute thus turns out to collapse into the supposition that he isn't time-traveling at all. Given any positive number *n*, an analogous argument shows that continuous forward time-travel at a rate of *n* minutes/minute collapses into forward time-travel at a rate of one minute/minute.

There is another, quite different problem with the concept of discontinuous time-travel. Let W_1 be a universe where (i) is true and (ii) is false; and let W_2 be an otherwise indistinguishable universe where (ii) is true and (i) is false. What happens in the one universe at any given time coincides with what happens in the other universe at that time. It follows that, even if (*per impossibile*) an occupant of the one universe knew exactly what happened in that universe at any given time, he would not, in virtue of that fact, have any legitimate reason for preferring (i) to (ii). *A fortiori*, no possible observation that one might have would provide one with any reason for believing oneself to be in W_1 as opposed to W_2. Since observation couldn't possibly provide one with such a reason, only logic could do so. But logic cannot do so, since (i) is logically false, as we've already seen.[432]

Similar points hold in connection with discontinuous backwards time-

travel. In the movie *Back to the Future*, young Marty McFly vanishes from the year 1985 and shows up in the year 1955. This means that, in the year 1985, the causal sequence that constitutes his existence is severed. Like everything else that persists for any length of time, Marty McFly is a causal sequence. It follows that, so far as it isn't incoherent, the statement:

> (1) In 1985, Marty McFly traveled backwards in time to the year 1955 and is otherwise unchanged.

collapses into:

> (2) Marty McFly was annihilated in 1985, and in 1955 there appears *ex nihilo* somebody who is numerically distinct from him but otherwise just like him.

Also, given obvious adaptations of what we said a moment ago, there is no conceivable event that, were it known, would provide more support for (1) than for (2).

Since, therefore, there couldn't possibly be any observation-based reason for preferring (1) to (2), the only possible reasons for doing so would be of a logical nature. This means that there are *no* reasons for doing so, given that (1) is incoherent. The concept of discontinuous backwards time-travel is therefore a broken one.

What about the idea of *continuous* backwards time-travel? For argument's sake, let's suppose that, during his continuous journey from the year 1985 to the year 1955, Marty McFly is right next to some stop-sign that was built in 1920 and stays in the same place for several centuries. That means that, at any given moment during that thirty year period, Marty McFly is no more time-traveling than the stop-sign. Each is just sitting there.

There are, of course, some important differences between Mary McFly and the stop-sign. First of all, whereas the stop-sign is crushed or burnt or otherwise destroyed in some pedestrian way, Marty McFly must simply vanish in the year 1985. (If he doesn't, then he's still hanging around in the year 1985, in which case he obviously hasn't gone back in time.) And, whereas the stop-sign's presence in the year 1955 is to be explained in the customary way, Marty McFly just popped up *ex nihilo.*

Bearing these points in mind, let W_3 be a universe in which:

(3) Marty McFly continuously traveled backwards in time from the year 1985 to the year 1955 but is otherwise unchanged.

is true, and in which (2) is false; and let W_4 be a universe that is exactly like W_3 *except* that, in it, (2) is true and (3) is false. Echoing what we said earlier, there is no conceivable event that, were it known, would enable one to determine whether one was in W_3 as opposed to W_4, or *vice versa.* Which, taken together with the fact that being annihilated in 1985 didn't exactly tighten young Marty's grip on existence, suggests that, other things being equal, (2) is the better way to model the observational data.

There is another, even deeper issue. For there to be backwards time-travel of any kind, there must be retroactive causation: some events must be the effects of *later* events. Marty McFly's appearing in 1955 must be an effect of some event occurring in 1985. But there are empirical and also strictly logical reasons for thinking backwards time-travel an impossibility.

If Relativity Theory (RT) is right, which we'll henceforth assume that it is, the statement "E_2 is an effect of E_1, but E_1 post-dates E_2" is tautologically false. According to RT, order in time is *defined* in terms of causation: anything caused by a given event *ipso facto* follows that event. Thus, the view that one could travel backwards in time is in the same category as the view that triangles could have five sides.

A well-known consequence of RT is that a given event may precede one event relative to one framework, be simultaneous with it relative to some second framework, and follow it relative to some third framework. Some conclude from this that backwards time-travel *is* possible.

This is not the right inference to draw. For reasons that aren't important here, the framework-relative nature of temporal order is a *consequence* of, among other things, the fact that any effect of an event *ipso facto* follows that event. In other words, it's a consequence of the fact that "E_2 precedes E_1 and is also an effect of it" is tautologically false.

Like any other scientific theory, RT is empirical. It's based on observation; and acceptance of it is warranted only as long as our observations are consistent with it. Nonetheless, the principle that causes *ipso facto* precede their effects is an analytic, not an empirical, truth. (We'll

henceforth refer to this principle “CT”—short for “causality and temporal order.”) So CT is true even if RT is false.

“But how can CT be analytic,” it will be asked, “given that RT is an observation-based theory?” Two reasons. (1) Not every statement composing an empirical (observation-based) theory is *itself* observation-based. Every theory comprises statements of pure logic (e.g., “no true statement entails the negation of anything entailed by any of its consequences”) that aren’t observation-based. (2) Not everything learned with the help of observation is learned *through* observation. If one learns some complex mathematical theorem by reading a proof of it, one learns that theorem *through* observations of ink-deposits; but it isn’t *by observing* such deposits that one learns it. The observations of ink-deposits merely occasion or “trigger” the purely logical ratiocinative activity that *does* lead to a genuine knowledge of that principle.[433] Somebody with 20/20 vision who has carefully studied those ink-deposits (i.e., who has studied those ink-deposits themselves, as opposed to their meanings) may completely and utterly fail to understand the corresponding proof; whereas somebody whose vision is just good enough to see some of the key lines of the proof may, on the basis only of those observations along with the thoughts they precipitated, understand it extremely well.

Also, somebody who has heard the proof stated but hasn’t read it may, on the basis only of those auditory observations along with the thoughts they precipitated, thoroughly understand the proof. Ink-deposits bear no physical resemblance to noises. Therefore, it isn’t by studying ink-deposits that one learns the proof.***[434]*** So given only that empirical discoveries led to our becoming aware of CT, it doesn’t follow that CT was itself discovered empirically.

And, in fact, CT couldn’t have been discovered empirically. Assuming it a certainty that Smith’s pushing the button caused the elevator to come, there is no conceivable event that, were it known, could warrant a revision of one’s subsequent judgment that the pushing of the button preceded the arrival of the elevator.

Einstein himself very clearly said that CT *was* not an empirical hypothesis. He also said (unfortunately for me) that it was a convention—a *definition* of “order in time”—and, therefore, not a substantive truth. I would suggest that, although he described CT as a definition, he chose this particular definition, as opposed to others, because he saw in it a substantive truth whose roots lie

in the very concept of what it is for one event to precede another.

Chapter 25 The Compatibilist Conception of Freedom

1.0 Determinism in relation to the debate between materialism and dualism

A set of events is deterministic if it obeys laws that, given the system's state at any given point in time, fix the system's state at all later points in time.[435] The physical world is a deterministic system. (In any case, it's very nearly so. And in this context, we'll assume that it is. Later, we'll drop the assumption.) This means that the world itself is a deterministic system *to the extent* that it's physical.

There are both physical and mental entities. Anything that involves thought or feeling is mental; anything that does not (but is spatiotemporal; i.e., is in space or time) is physical. So beliefs, intentions, and regrets are mental; and rocks, trees, and explosions are physical.

It is an open and delicate question how the mental and physical are related. (See Chapter 22 for a discussion of this matter.) Some hold that mental entities are *identical* with certain physical entities (presumably neural events or structures). This position is known as ***materialism.*** Others deny this, saying that the mental is non-physical. This position is known as ***dualism.***

If materialism is right, then the mental world is just part of the physical world. So if *E* is some experience that you either have had, are having, or will have, then given the way the world was 10 million years ago, the state of the universe 10 million years ago made *E*'s occurrence inevitable. Whatever you decide to do today is an inevitable consequence of the way the world was a million years ago.

Supposing that the physical world is deterministic, and supposing also that materialism is right, it's hard to see how there could be free will. If the state of the universe a million years ago completely settled what choices I would make today, then presumably I couldn't make any choices other than those that I *do* make. And it would seem that, under those circumstances, I don't have anything resembling "freedom of choice." In a moment, we will find that some case can be made that one may have *a* kind of freedom in a deterministic world. But let's hold off on that for a moment and talk right now about the relationship between *in*determinism and freedom.

2.0 Free will in an indeterministic world?

Because of the apparent irreconcilability of materialism with the presumption that we are free, some believers in free will accept dualism, while others accept materialism but reject determinism.[436] Let's consider both options, starting with the first.

First of all, supposing that there is indeterminism in the world, it doesn't necessarily mean that our mental processes aren't determined. If the atomic processes making up some rock happen to be indeterministic, that doesn't mean that your brain or mind aren't deterministic. So the indeterminism must be in the right place.

What is that place? Where must the indeterminism be if our choices are to be made freely? The indeterminism must relate to the processes involved in making choices. For, if the mechanisms of which our choices are results are deterministic, then our decisions are as determined as anything else. Thus, a free person's choices are undetermined.

But a choice that isn't predetermined by anything isn't predetermined by any facts about one's personality, and a choice of the latter kind is not one that we would have any inclination to regard as "free." A short story may illustrate my meaning. Smith is walking along. All of a sudden, he decides to become a pilot. This decision has no basis in his personality. He never liked flying. He already has a career that he enjoys and he depends on. So it would be inconvenient for him to become a pilot, and he would derive no pleasure from it. He knows all of this. Nonetheless, he makes this choice. It just comes out of the blue.

Was this choice made *freely?* It doesn't seem like it. It actually seems rather *compulsive*—like something that just overcame Smith. Suppose that "choices" similar to the one just described were a frequent occurrence in Smith's psychological life. Would we say that Smith was "free"? No. We'd describe him as having a kind of disorder. We'd describe the fact that he was visited by these strange compulsions as indicative of a lack of psychological integrity on his part; and see his submission to these forces as showing a *lack* of freedom on his part. Smith is a puppet dancing at the end of strings pulled by psychopathology.

This suggests that, in order to be made "freely," a choice has to be rooted in facts about one's personality, and must express one's values, desires, beliefs, and previous intentions. A choice that lacks these roots seems to involve a

disintegration of the person's psyche and, paradoxical though this may seem, a consequent diminution of the subject's freedom. So even if there is indeterminism, it doesn't help people be free; indeterminism actually seems to hurt our chances of being free.

But to the extent that a choice *is* appropriately rooted in a person's personality, that choice is *not* the result of indeterminism but, on the contrary, of a deterministic structure. And it's hard to see how, under those circumstances, we could be free. So we're damned if we do and damned if we don't!

3.0 Free will in a dualistic world?

Now let's consider the view that, although the physical world is deterministic, the mental is not physical. First of all, even if the mental is not physical, there is considerable evidence that the mental sphere is deterministic. The decisions that one makes are clearly a product of one's pre-existing mental condition.[437]

We must also take into account that, at least according to some viewpoints, the choices we make have extremely deep roots in our unconscious mental lives. Analysis tends to suggest that decisions that seem, at the level of consciousness, to be "random" are, in fact, strictly determined by unconscious forces. In any case, this is one point of view, and it is one for which there is some corroboration.

But suppose, if only for argument's sake, that your choices were not fixed by the state of the universe prior to their occurrence. For the reasons given earlier, that wouldn't make those choices be free. On the contrary, it would make them more like fits or seizures than like considered, freely made judgments.

A story similar to the one just told may help make this clear. You are a student at State University. You have an A-average. All of a sudden, you choose to drop out of college and become a drug dealer. Your choice was not caused to occur by anything; the state of your personality prior to the choice had nothing to do with your making it, and neither did the state of the universe. It just happened. So, in the middle of a test, you throw down your pen, bolt out of the room, and get to work as a drug dealer.

Does this sound like a freely made choice? No. It sounds more like a compulsion. It's as though you were "taken over" by another person and lost control of your body. Again, we see that indeterminism doesn't help freedom and actually undermines it.

4.0 The Compatibilist Conception of Freedom

The problem is that, as we saw, determinism *also* seems to threaten freedom. Obviously you have no control over the past. If determinism is right, then what happened in the past completely determines exactly what's happening in the present and, in particular, completely fixes what choices you're making now. So something over which you have no control is completely responsible for the choices that you're making. It's hard to see how, under these circumstances, you could be said to have free will.

Nonetheless, some hold that there *can* be free will in a deterministic world. Such people are known as ***compatibilists.*** Those who deny it are ***incompatibilists.***

4.1 The crude form of compatibilism

There are ultimately two different forms of compatibilism. One of these is very straightforward; the other is quite sophisticated.

According to the straightforward version, which we'll discuss in this section, you are free to the extent that you can do what you want, and you are unfree to the extent that you cannot. If you want to leave your room and you can, you are free. (In any case, you're free *in that respect.* You may be unfree in other respects; e.g., once you leave the room, you may want to leave the house you're in but not be able to.) According to the other version of compatibilism, which we'll discuss in Section 4.2, being able to do what you want, though necessary for freedom, isn't sufficient for it. It is necessary also that your choices be adequately rooted in who you are as a person.[438]

4.1.1 G.E. Moore’s compatibilism

According to G.E. Moore (1873–1958),[439] an act is performed freely if the agent *would* have done something else *if* he had chosen to do so. An act is free if the agent “could have done otherwise.”

The idea is pretty straightforward. Although I’m considering striking Harry, I choose not to. But, immediately upon making that choice, an extremely strong person grabs my arm and forces me to hit Harry. My act was clearly unfree. But why? Because, says Moore, even though I chose to do otherwise, I ended up striking Harry. Now suppose that I *voluntarily* strike Harry. I do it, not because somebody forces me to, but because I don’t like Harry and want him to suffer. In this case, my act *is* free. But why? Because, says Moore, if I hadn’t chosen to strike Harry, I wouldn’t have done so. Thus, my striking Harry is voluntary, Moore says, iff I wouldn’t have done so had I chosen not to do so.

Moore’s position is a form of compatibilism. Free acts aren’t those that are uncaused; they’re those that are caused in a certain way—they’re those that are caused by decisions on the part of the people (or animals) who commit them.

4.1.2 Chisholm’s objection to Moore’s analysis

In *Human Freedom and the Self,* Roderick Chisholm (1916–1999) makes a compelling objection to this analysis. Let’s suppose that, if I had chosen not to strike Harry, I would indeed have refrained from doing so. Unless I was at liberty to *not* make[440] that choice, it’s hard to see how my doing so was free. If every choice I make is inevitable, then it’s irrelevant that, had I made different choices, I would have acted differently. For I *couldn’t* have made those choices under those circumstances, and I therefore *couldn’t* have acted in those ways. And if I couldn’t have acted in those ways, then I couldn’t have done anything other than what I did, making it hard to see how, in acting as I did, I was acting freely.

4.1.3 A Lockean counter-objection

Though powerful, Chisholm's point involves a serious confusion, which may undermine it. We won't identify that confusion until Section 4.1.5, since we haven’t yet covered some of the material that must be absorbed to understand it.

Right now we will concentrate on a point made by John Locke (1632–1704), who held a view similar to Moore’s, that either refutes, or comes close to refuting, Chisholm’s analysis.[441] In Locke’s view, freedom is a property of *actions.* It makes no sense to speak of anything other than actions as being free. Only things that one can *intend* to do are done freely. I can *intend* to do 20 push-ups, and that’s why I can do it freely. But I can’t *intend* to make a certain choice. What would it be to *intend* to make a certain choice (e.g., to intend to do 20 push-ups)? If I really *intend* to choose to do 20 push-ups, then, for that very reason, I *have* chosen to do 20 push-ups. Second-order intentions (intentions to intend) collapse into first-order intentions. Thus, so far as it isn’t a pleonasm, the expression “intentional choice” is meaningless.

4.1.4 A Lockean counter-objection (continued)

Locke has surely identified one kind of freedom or---what, as we'll now see, is probably more accurate---an essential *component* of one kind of freedom. But the following scenario suggests:

> (i) that Locke's analysis, though certainly a *part* of the truth, is not the whole of it;
>
> (ii) that, although Locke's analysis may well partially undermine Chisholm's analysis, it doesn't do so completely;

and, finally,

> (iii) that Chisholm's has shown that the concept associated by the lay-person with the term "free" is incoherent and, moreover, that the same is true, though not for the same reasons, with Locke's attempt to remove the incoherencies from that concept (while leaving it otherwise unchanged).

A story will clarify these points.

On Monday, you undergo hypnosis. You do so because you believe that it will help you kick your addiction to nicotine. While under hypnosis, the hypnotist, who is a prankster and loves to abuse the power he has over his subjects, tells you that, on Wednesday, you should *choose* to put on a tin-foil hat, go into your boss's office, and tell him that aliens are about take over the Earth. On Wednesday, you choose to do these things. So you *voluntarily* put on a tin-foil hat, etc.

Your act was voluntary. You chose to do it. But was it free? My opinion: there was *a* sense in which it was free, but there was also an important sense in which it was not free. You are not, by virtue of choosing to engage in this odd behavior, in the same category as somebody who wants to walk to his car but is physically restrained. But, for reasons that are not immediately obvious but that we will now try to elucidate, your behavior is either unfree or free only in some attenuated sense.

One way of interpreting what Chisholm is saying is as follows[442]:

(CR) There is only one significant difference between, on the one hand, your voluntary decision to don a tin-foil hat, etc. and, on the other hand, a 'normal' decision (e.g., a decision to play tennis made by somebody whose choice isn't the result of hypnotic suggestion or any other comparable event). The difference in question is that, whereas your decision was implanted in you by some sentient being, the tennis-player's decision was implanted in him by impersonal forces. But does that really make a difference? To be sure, your decision (to put on a tin-foil hat, etc.) is the result of malfeasance on somebody's part, whereas the tennis-player's decision is not. So, from a moral perspective, there are important differences between the two situations. But, when evaluating whether a person's decision to do *X* is free, what is relevant is whether that person had control over the forces leading to that decision; it's irrelevant what sort of morality, if any, those forces embodied.

An adaptation of the story told a moment ago (about the hypnotist) makes this clear. One day, while you are snoozing under a coconut tree, a coconut falls on your head. The resulting brain-disturbances lead to your deciding to put on a tin-foil hat, go into your boss's office, etc. Surely your decision to do this---though voluntary, and though not the product of malfeasance on anyone's part---is unfree or, at best, is free only in some reduced sense. So Locke's thesis ("an act is free iff it's voluntary") isn't correct and must either be rejected or at least heavily qualified.

4.1.5 An oversight on Chisholm's part

Partly for the reasons just given, and partly for reasons to be stated later in this chapter, I believe that Locke's analysis is inadequate. To that extent, I must agree with Chisholm. But Chisholm's analysis involves a subtle but important confusion. Because it involves this confusion, Chisholm's analysis fails to show that actions cannot be free. (However, despite its involving this confusion, Chisholm's analysis---or, rather, the extension of it put forth in the previous section---does successfully undermine Locke's position.)

Let *t* be some instant in the year 50,000 B.C., and let *t** some instant in the 21st century. Suppose that, at *t**, Smith decides to go for a jog. Bearing this in mind, consider the following two statements:

> (1) The state of the universe at time *t* compelled Smith to decide to go for a jog at time *t**.
>
> (2) The state of the universe at time *t* made it inevitable that, approximately 52,000 years later, Smith would come into existence a few years prior to *t** and, because of the combined effects of his pre-*t** experiences and his native constitution, would decide at time *t** to go for a jog.

(1) entails that, given the state of the universe at *t*, it was nomically impossible (i.e., prohibited by natural law) for Smith *not* to decide to go for a jog at *t**; and (2) also entails this. But there is a difference between (1) and (2). (1) says that the state of the universe at *t* compelled Smith to make a certain decision. (2) doesn't say this. It says that it was inevitable that he would decide at *t** to for a jog. But that's compatible with his *not* being compelled at *t** to decide to go for a jog. It's inevitable that I will choose to have a meal in the next 12 hours; but that doesn't mean that my decision to eat isn't free. Rational people freely but inevitably choose to act in certain ways.

Here's a case of Smith's being *compelled* to behave in a certain way. A very strong person takes hold of Smith's arms and, despite Smith's valiant resistance, makes Smith hit himself with his own hands (or, what is probably a more accurate way of putting it, hits Smith with Smith's own hands). Why

is it that, in this case, Smith's behavior is the result of compulsion? Because the immediate causes of that behavior were not Smith-internal. The forces that led to Smith's arm-movements were not constitutive of Smith; they failed to be Smith-constitutive even in the attenuated sense in which an impulse is constitutive of a person of a person who regards that impulse as having no value, having thus made every effort to dissociate himself from it, and to which he therefore gives no weight when making decisions.

Even though (2) is to the effect that it is compulsory that Smith come into existence and, in due course, do such and such, (2) is quite consistent with *Smith's* never being compelled in any way. For (2) is consistent with the proposition that the thing that is being compelled is the *universe*, not Smith, and that what the universe is being compelled to do is to produce a being, namely Smith, who will decide at *t** to go jogging, and whose decision to do so, having for its proximal cause a conglomerate of Smith-constitutive forces, is not compelled---at least not in the sense in which, in the other story, Smith was compelled to hit himself and, indeed, not in any ordinary sense of the word "compelled."

A continuation of our story will help us move forward. The state of the universe at *t* makes it necessary, as a matter of natural law, that Smith come into existence, that he have thus and such thoughts, feelings, beliefs, etc. and that, at time *t**, he decide to go jogging. But the state of the universe makes it inevitable that Smith's decision to go jogging be the result of forces that *make up Smith.* Obviously Smith's decision will reflect his beliefs concerning the external world; and those believes are obviously, in part, the results of events (collisions of photons with Smith's eyes, etc.) over which Smith has no control. But these Smith-*non*-constitutive events caused Smith to choose to go jogging only *by way of* Smith-constitutive forces---by way of Smith's preferences acting in conjunction with Smith's beliefs. So, ultimately, nothing that wasn't Smith-constitutive made Smith do anything. The only thing that made Smith go jogging was Smith himself. So, even though natural law made it inevitable that Smith would come into existence and, at time *t**, decide to go jogging, the way in which natural law made that inevitable didn't involve Smith's being compelled. It involved the world's being compelled to produce a person (Smith) whose own self was the sole (proximal) cause of his choice and who therefore *freely* went jogging. So (2) is compatible with the following proposition:

(2*) Given the state of the universe at *t*, it was necessary, for reasons of natural law, that the *universe* be compelled to produce a person (Smith) who, at time *t**, *freely* went jogging.

If *you* were the sole immediate cause of some action of yours, then you acted freely. It may be that your existence, along with everything you think and feel (etc.), is the inevitable result of how the world was before you were born. But, by virtue of producing you, the world was not compelling *you* to do (or be) anything. There was no you to compel. (In producing somebody, you are not *compelling* them to exist. In conceiving a child, you are not *compelling* that child to exist. The child isn't yet in existence and therefore isn't there to be compelled. What is being compelled is the coming into existence of that child. The universe is being compelled, by your carnal acts, to create junior. But junior himself isn't in any way being compelled, at least not by virtue of being conceived, since what doesn't exist can't be compelled.)

And once you are in existence, your actions are free if their causes are forces that are constitutive of you. For so far as your deeds result from such forces, they result from *you*: *you* are the author of those actions. And even though the state of the universe prior to your birth made it inevitable that you would exist to perform those actions, it did not (setting aside cases where forces not constitutive of you; e.g., a bully's arm-twisting, are the proximal causes of your behavior) compel *you* to perform those actions. Rather, the state of the universe prior to your birth compelled events to produce a person (you) who (setting aside bully-caused behaviors, etc.) would freely perform various acts.

4.1.5 Frankfurt's refutation of Moore: why "*x* acted freely" ≠ "*x* could have done otherwise"

Harry Frankfurt (1929–) decisively demonstrated the falsity of Moore's position by identifying scenarios in which somebody performs some act freely even though that person could *not* have abstained from performing that act.[443] Here is such a scenario:

> (SM)[444] Mr. Smith intends to kill Mr. Jones. Like Mr. Smith, Mrs. Green also wants Mr. Jones dead. But Smith doesn't want to go to jail for murder. Without Smith's knowing it, Mrs. Green installs a microchip in his cranium. (We needn't worry about how exactly she does this or how exactly the microchip works.) By means of this microchip, she can control Smith's actions—she can make him do whatever she wants. She knows that Smith is intent on killing Jones. But if Smith should get cold feet, Mrs. Green will use the microchip to force Smith to go through with it. Smith doesn't know about the microchip; nor does he even know of Green's existence. Smith kills Jones—he doesn't chicken out at the last minute. And he does so freely (at least insofar as anyone does anything freely—he is not in the grips of any delusions, or compulsions, etc.).

Thus, Smith freely kills Jones, even though Smith could not have done otherwise. For, had he decided at the last minute *not* to kill Jones, Green would have forced him to.

Hence, contrary to what Moore alleges, "*x* performs act *A* freely" is *not* equivalent with "*x* does *A* and, had *x* chosen *not* to do *A*, *x* would not have done *A*."

4.1.6 The irrelevance of Frankfurt's refutation of the counterfactual analysis of freedom

But even though, interpreted narrowly, Moore's analysis is false, it's really just a misguided way of stating the principle, which Frankfurt's argument does nothing to undermine, that a free act is one that is choice-driven. Moore was led to equate "*x* acted freely" with "*x* could have done otherwise" *not* by a misconception concerning the nature of freedom, but by his acceptance of a certain analysis of causality—the so-called "***counterfactual analysis of causality***" (CFA). According to CFA, *x* caused *y* to occur just in case if *x* hadn't happened, *y* wouldn't have happened. "Smith's pushing the button caused the elevator to come" means the same thing as "if Smith hadn't pushed the button, the elevator wouldn't have come."

CFA is wrong. This is because there is causal redundancy in the world. I push the elevator button, causing the elevator to come; but if I hadn't pushed it, one of the other 20 people waiting for it would have done so, and the elevator would have come.[445]

But a close cousin of CFA is correct, namely:

> (CFA*) *x* caused *y* to occur just in case, if *x* hadn't happened, then ***ceteris paribus*** *y* wouldn't have happened.

Similarly, even though, for the reasons just given:

> (PAP)[446] "*x* did *A* freely if and only if *x* did *A* and, further, had *x* chosen not to do *A*, *x* wouldn't have done *A*,"

is false, nonetheless:

> (PAP*) "*x* did *A* freely if and only if *x* did *A* and, further, had *x* chosen not to do *A*, then *ceteris paribus x* wouldn't have done *A*,"

is correct.

It's well known that, if they are to come out correct, counterfactuals

typically have to be hedged with *ceteris paribus* clauses. Even though:

> (JFK) "If JFK hadn't been assassinated, he would have been re-elected,"

is a plausible counterfactual, it isn't necessarily true. What if, sometime after Nov. 22, 1963, but before the '64 elections, JFK died of natural causes or was disgraced by some scandal? In that case, he would not have been re-elected. So the right counterfactual is:

> (JFK*) "If JFK hadn't been assassinated, then *ceteris paribus* he would have been re-elected."

In other words, if JFK hadn't been assassinated, but *otherwise* things were the same, so that JFK was healthy and popular, etc., *then* he would have been re-elected.[447]

Notice that JFK* is really a causal claim, since it's just another way of saying:

> (JFK#) "JFK's being assassinated caused an otherwise re-election-bound JFK to *not* be re-elected."

As a rule, counterfactuals are causal claims.[448] Thus, "Smith killed Jones freely" is true just in case (i) Smith killed Jones and (ii) if Smith had chosen *not* to kill Jones, then *ceteris paribus* Smith would have not killed Jones.

Basically, Smith killed Jones freely if his doing so expressed his decision to do so, as opposed to some other force. In general:

> (FA)[449] an act is free if it expresses one's decision to act that way.

This is clearly the right analysis of *a* kind of freedom or---what may be a more accurate way of putting it---a certain *aspect* of freedom.

4.1.7 Freedom not a property of choices

If FA is right, *acts* are free, not *choices*. This is what Locke held. Locke was the first advocate of FA, and he was an exceptionally effective one. One chooses to act, Locke said; one doesn't choose to choose. It makes no sense to suppose that one chooses to choose. For if one always chose what to choose, Locke points out, one would have to make infinitely many choices in order to make a single one, and one therefore couldn't make any choices at all. In order to choose to open the door, one would have to choose to choose to open the door; and in order to choose to choose to open the door, one would have to choose to choose to choose to open the door; and so on.

So freedom is freedom to act, not freedom to choose. As Locke puts it, freedom is a relationship between choice and act, and not between subject and choice.

It follows that, so far as one's freedom is concerned, it is irrelevant whether one's choices are caused or not. Since freedom is a relationship between choices and acts, not between choices and their own antecedents, it doesn't matter whether choices are caused. Nor, therefore, does it matter *how* they're caused.

In fact, if FA is right, determinism is not only compatible with freedom, but is a prerequisite for it. Wanting to make the elevator come to your floor, you intend to push the button. If a degenerative neurological condition has undermined the integrity of the mechanisms linking your brain to your hand, thereby making them indeterministic, your intention to move your hand may be more likely to fail than to succeed. And, supposing that you do manage to push the button, if the mechanisms linking the button to the elevator are indeterministic, your pushing the button may be more likely than not to fail to get the elevator to come. To the extent that the mechanisms involved in the implementation of our choices are indeterministic, we are in the grim position of somebody who has Parkinson's. The spirit is willing, but the flesh can't be counted on to obey, and neither can the relevant mechanisms in the external world.

5.0 Frankfurt's analysis of freedom[450]

According to Locke, you are free to the extent that you can do what you want to do, and you are unfree to the extent that you cannot. This is obviously

a kind of freedom—a very important kind.

But it isn't the only kind. Contrary to what Locke says, there is *a* sense in which choices can be free or unfree. Harry Frankfurt made this clear.[451] A story will help us begin to understand Frankfurt's insightful analysis of freedom.

Jones is a great writer who is addicted to heroin. Because his addiction is only moderate, he is able to function perfectly well even when he is heroin-deprived. (When heroin-deprived, he doesn't feel good, but he is in no way incapacitated.) He knows that, if he were to shoot up right now, the subsequent lethargy would make it impossible for him to write a single word for the rest of the day. He also knows that, if he doesn't write 10 pages today, he will lose his book contract. Jones does not want to lose the contract. Plus, he wants to grow as a writer, which he won't do, at least not today, if he shoots up. He thus has a desire to *not* take the heroin—even though he simultaneously has a desire *to* take the heroin.

Jones' desire to take heroin is a first-order desire. Jones' desire *not* to act on this desire is a ***second-order desire.*** This is because Jones' desire not to take heroin is a desire as to what to do with another desire; it is a desire *not* to cave in to a desire to take heroin.

Suppose that Jones caves in and he takes the heroin. In that case, Jones is a slave to his desires. His will has been compromised. He is not acting freely.

But *now* suppose that Jones does *not* cave in. He does the right thing: instead of taking the heroin, he endures the discomfort of a heroin-free day and manages to get his work done. In *this* context, Jones is not a slave to his desires. His will has not been bent. His conduct expresses a "free will."

In light of considerations such as these, Frankfurt says that somebody has a free will to the extent that his conduct aligns with his *second*-order desires.

But here a question arises. Why are second-order desires so special? Why is it that freedom consists in one's doing the bidding of one's *second*-order desires—as opposed to one's third-order, or fourth-order, or first-order desires? What, in this context, is so special about the number two?

When the heroin-addicted writer decides *not* to take heroin, what is driving him is a rational appraisal of his situation. But when the heroin-addicted writer gives in to his addiction, it is *not* his rationality (or his morality) that is driving him.

Before we can close this argument, we must address Frankfurt's discussion of the concept of a "person." By a "person," Frankfurt does not

mean a member of the species *homo sapiens.* He means anything that has the same basic psychological architecture that we have. What is so special about that structure? First of all, we are self-conscious. Second, we are rational. Third, we are rational not only about how to attain our objectives, *but also about what those objectives should be.* What distinguishes persons from non-persons is that the former, but not the latter, subject their own desires (and, more generally, their own mental states) to critical scrutiny.

Non-persons are often very clever when it comes to figuring out how to attain their objectives; for example, beavers show a lot of intelligence when it comes to building dams. But they are not smart about whether they should be building dams in the first place. The question doesn't arise. They have the desire, and they act on it (or they die trying). In general, persons do, whereas non-persons do not, subject their own desires to rational scrutiny. Non-persons do not ask themselves: "What would the consequences of acting on this desire be? And *should* I act on this desire?"

We persons, on the other hand, *do* subject our objectives to rational assessment. And this is obviously made possible by the fact that we are self-conscious. If I am overtaken by a desire to build a dam, I will ask myself: "how would my acting on this desire help my overall interests?"

We can now close the argument. Whenever someone is driven by a rational assessment of their own desires, they are driven by a desire about a desire. Thus, one is rational only to the extent that it is one's second-order desires that drive one's conduct. There is a sense in which rational conduct is free and in which irrational conduct is not.

To make it clear what this sense is, we must first distinguish between intelligent conduct and rational conduct. An irrational act can involve an enormous amount of intelligence. Smith is a brilliant inventor. He tries to impress his wife by building a giant robot that sings romantic songs in a metallic voice. This alienates her and she divorces him. Moreover, the financial and personal resources he sunk into this project have left him destitute. Smith's act might have been about as intelligent as an act can be. But it was not a rational act. For intelligently carried out though the act was, the immediate object of that act (namely, the construction of a robot that would sing love songs to his wife so that he'd be spared the trouble of doing so) was not intelligently arrived and failed to embody consideration of relevant facts that it was well within Smith's ability to consider.

A rational act is one that embodies due regard for one's various interests.

An irrational act is one that does not. An intelligent but impulsive act (e.g., Smith's speedily building this robot to allay his rapidly growing anxieties about his marital situation, all of which are projective but which become legitimate as a result of this endeavor on his part) fails to embody due regard for one's own interests. (Smith isn't taking into account the consequences of his actions; he is misdeploying his vast stores of intelligence. And, granting that there are different kinds of intelligence, he obviously has whatever kind of intelligence one needs to have to know that (supposing for argument's sake that there's no chance that this project would bring him to the attention of any potential benefactors) this project is bound to land him on Skid Row.) So far as one is driven by first order desires, one is neglecting to gratify one's second order desires. So, given that second order desires are clearly more closely hewed to one's middle- and long-term welfare---or, if it isn't one's own welfare that one cares about, one's middle- and long-term situations---it is clearly more rational to act on one's second than on one's first order desires. Also, to make a point we'll develop in Chapter 26, it's often rational to perform actions in which, were one's first-order desires allowed to discharge themselves, they'd be expressed. *Too* much self-restraint is dysfunctional and, in addition to being unpleasurable, seals off possible well springs of insight and creativity. But, setting aside cases of psychopathology, when a person performs one of these enjoyment-oriented actions, he isn't acting on an impulse; he's acting on his *approval* of an impulse. In other words, he's considered the impulse, determined that it lines up with second order desires (or, what I believe to be different, his values), and, on that basis, lifted the initial inhibition on his impulse. So, while the resulting action may be indistinguishable from what he would have done had the impulse itself been the proximal cause of his deeds, his psychological condition, given that his action was *not* directly prompted by that impulse, is radically different from what it would have been had that action been thus prompted.[452]

Having distinguished between rational conduct and intelligent, we may proceed with our discussion of Frankfurt’s analysis of freedom.

5.01 Frankfurt's analysis of freedom (continued)

There's more to being free than being able to do what you want to do. Your desires must reflect who you are. If they don't, those desires *belong* to you but aren't *of* you. And in that case, you are alienated from your own mental states, in much the way that a person with Parkinson's is alienated from his body.

But, in a way, the alienation is even worse. Involuntary behavior is unintentional behavior. Intentional behavior is voluntary behavior. Voluntary behavior is desire-driven behavior. (If you don't *want* to raise your arm, but it goes up anyway, you didn't intend for it to happen.) Epileptic fits are not *actions.* Unintended behaviors are not *actions.* Unlike actions, they're *undergone,* not *done.* So estrangement from one's own actions amounts to estrangement from all of one's own desires; and it's hard to imagine a condition less free than that.

5.1 Frankfurt's emendation of this analysis

Another story will help expose a problem with the analysis of freedom just described, and it will also clarify the raison d'être for Frankfurt's soon to be discussed emendations of it.

Smith's first-order desires are seriously out of alignment with his second. He values writing poetry; he's good at it; and he enjoys it about as much as it's humanly possible to enjoy it. But he has a voracious craving for drugs. His craving for drugs, especially cocaine, is unusually strong; and his desire to write poetry, though intense, is not as intense as his desire to do cocaine.

Smith's desire to do cocaine is out of alignment with many of his values (second-order desires). Were he to give into this desire, he couldn't write poetry; nor could he maintain friendships or play sports or, with only a few exceptions, do any of the other things he values. For this reason, he doesn't give in. But given how intense his craving is, every day is a struggle for Smith. If, as Frankfurt suggests, one is free to the extent that one acts on one's second-order desires, Smith is as free as a person can be. But given how hard life is for Smith, it's counterintuitive to suppose that he's maximally free. To be sure he has *a kind of* freedom, and he has it in abundance. But there's some other sort of freedom that he doesn't have.

Frankfurt is aware of this problem, and he revises his view to deal with it. According to his revised view, if an act is to be free, it is *necessary* that it be driven by a second-order desire, but it isn't *sufficient.* Another condition must be met: one's first-order desires must be consistent with the second-order desires that are driving that act.

Here's the idea. Person *X* has no self-control. He binges on drugs and food, and he hates himself for doing this; but he can't stop—he's too weak and his urges are too strong. His life has no meaning and he knows it; but he will not, or cannot, do anything about it. Unlike *X*, Person *Y* has a huge amount of self-control and never caves into temptation. But everything is a struggle, since he's always fighting temptation, like a drowning man trying to stay afloat. (*Y* is like Smith.) Person *Z* cannot be categorized with *X* or with *Y*. *Z* always does what he values, which is composing music. And what he values is what he *wants.* He wants what he wants to want. Because he's doing what he wants, he has the freedom that *X* has that *Y* lacks. Because he's doing what he wants to want, he has the kind of freedom that *Y* has that *X* lacks. And since, of the three, he's the only one who has both kinds of freedom,

he’s obviously the most free of them all.

Thus, one is free insofar as two conditions are met: (i) one acts on one’s second-order desires, as opposed to one’s first-order desires; and (ii) one’s second-order desires are in alignment with one’s first-order desires.

To see the merits, and also the shortcomings, of Frankfurt’s brilliant analysis, we must take a moment to say how that analysis relates to compatibilism.

5.2 "*x* is free"= "*x* does what *x* wants to *provided* that what he wants to do is adequately rooted in his personality structure"

We've considered two analyses of freedom: (i) "*x* performed act *A* freely" = "it is because *x* chose to do *A* that he did it," and (ii) "performed *A* freely'' = "*A* reflects who *x* really is." These two views can be combined into the view that (iii) "*x* performed act *A* freely" = "it is because *x* chose to *A* that he did it, *provided* that this choice of *x*'s reflects who *x* really is."

This analysis, which is basically Frankfurt's, embodies a profound insight into the nature of human freedom. We'll see in Chapter 26 that, although it's a step in the right direction, it isn't correct: for it is possible to warp one's higher-order desires to suit one's first-order desires and, thus, to make one's higher-order desires be mere vehicles for first-order desires.

But right now, I'd like to concentrate on an apparent problem with Frankfurt's analysis that, although, I suspect, some would be inclined to think otherwise, isn't an *actual* problem with it:

> (FD)[453] Smith has given himself over to drug addiction. He spends days doing drugs. (He has unlimited supplies of money and never has trouble getting them.) Smith has many talents—he's a great water-skier, a great short-story writer, etc., but, because of his addiction, these gifts of his lie fallow.
>
> Green, on the other hand, hasn't given himself over to addiction. He spends his days doing things that augment who he is. Green can't indulge every craving that he has. Unlike Smith, he doesn't have the money to do so.
>
> You (JMK/Frankfurt) would say that Green is freer than Smith, and your reason would be that, even though he can't gratify as many desires as Smith can, the desires that Green does gratify have the right roots, whereas those that Smith gratifies do not.
>
> I grant that Smith's condition is worse than Green's, and also that Smith is less free than Green. But, contrary to what you say, that *is* because, when it comes to desire gratification, Green's batting average is better than Smith's. There are a lot of aspirations that Smith has that will be thwarted as long as he's imprisoned by addiction. Smith wants to do cocaine *more* than he

wants to be a writer, but it doesn't follow that he *no longer* wants to be a writer. He could very well want both.[454] Let's suppose he does. In that case, a very strong desire of his is being frustrated. It's true that *he's the one who's frustrating it,* but that doesn't mean it isn't frustrated.

Smith will inevitably have many other desires that his cocaine addiction dooms to frustration. In his pre-addiction days, we may suppose, Smith enjoyed many things that he no longer enjoys. We can assume that he *desired* to do each of these things, and we can also assume that he *still* desires them, even though he obviously doesn't desire them as strongly as he desires cocaine. Many desires are rooted in personality structures. A desire to play the piano or read Nietzsche doesn't just overtake one; it's rooted in enduring structural facts about one's personality. And as long as the personality structure is there, one's inability to express it in the form of action is, practically by definition, a form of frustration. Smith's personality structure may still support those desires. (And almost certainly will. Addiction doesn't turn people into complete vegetables, at least not right away, and usually not ever.) It could be that Smith *thinks* it would be greater. But it's not so clear that it would in fact be greater. The phenomenological expression of that frustration would be more intense. But that doesn't necessarily mean that the desire in question is itself more intense. Desires, even short-lived ones, are structures; they aren't *feelings.* And the feelings to which a desire gives rise don't necessarily make it clear how intense that desire really is.

FD is close to the truth, but it falls just short. Another story shows why. Aaronson is a coward who has no self-confidence. So even though he wants to write novels and is extremely talented at it, he doesn't do so. It isn't safe enough. It's too bohemian for his stodgy, blinkered parents and for his equally stodgy friends (who he doesn't even like—he just pretends to like them, since he's supposed to). To make a living, Aaronson sells insurance, which he hates. His life is a waking death.

In living as others want him to, Smith is surrendering his freedom *and* he is also dooming many desires of his to frustration. But those two failings, though

related, are not identical. Aaronson is suppressing an important side of himself. The personality structures that he is suppressing *give rise* to many ardent desires, but they aren't *identical* with such desires. His losing his freedom consists in his suppressing these structures. The desire frustration that follows is an *effect* of this loss of freedom and isn't *identical* with it.

So Aaronson's loss of freedom isn't identical with his being unable to do what he wants; his being unable to do what he wants is a by-product of his loss of freedom. To be sure, his inability to do what he wants is *itself* a loss of freedom. But, important though this second kind of freedom is, it's distinct from the first kind and a derivative of it. So being free consists primarily in being oneself, and secondarily in being able to do what one wants. And this is obviously close to what Frankfurt is saying, if it doesn't coincide with it.

5.4. Frankfurt's analysis systemically flawed

Many analyses that philosophers put forth seem not to embody genuine attempts to find the truth. Frankfurt's analysis of freedom *does* embody such an attempt. And that attempt is *almost* successful.

But it falls short. In the next chapter, we'll see why. Here's an outline of what we'll find. Acts that are driven by one's Higher-order desires need not be free; for people can, and often do, warp their higher-order desires, so as to make them fit their first-order desires. Higher-order desires, in other words, can become vehicles for first-order desires. A lazy, weak person can warp his higher-order desires so that they merely rubber-stamp the craven first-order desires of which his actions are compulsive expressions.

We will find that truly free acts are those that are driven, not by higher-order desires, but by *values*. We will find that there is no agency of any kind where there are no selves, and we will find that, where there are no values, there are no selves, though there may be *minds*. I will leave it at that for now, since the entirety of the next chapter will be spent decrypting these remarks. I'll end the present chapter by making some points about psychopathology that bear out what we'll say in the next chapter and also make it clear to what a large extent what we're going to say is derivative of what Frankfurt already has said.

6.0 Ego-dystonic vs. ego-syntonic

To be free, we have argued, one must identify with one's desires—gratifying them isn't enough. Psychologists refer to mental contents that one identifies with as ego-*syntonic* and to those that one doesn't identify with as ego-*dystonic*. An understanding of this distinction may clarify why determinism is necessary for freedom and why indeterminism threatens it.

Some people who suffer from psychological disturbances *identify* with their symptoms. They don't see their symptoms as symptoms. For example, the acute schizophrenic who sees goblins dancing on his bed thinks that he's seeing things as they are. To his mind, it's other people who are wrong: they can't see what's there; he can. The same holds of the paranoiac who thinks her phone conversations are being monitored by the President.

These are ego-syntonic disorders. The viewpoint embodied in the patient's symptoms coincides with the viewpoint of the patient himself. The latter doesn't regard the former as an "alien intruder."

But not all psychological disturbances are like this. The classic example is obsessive-compulsive disorder (OCD).[455] Those afflicted with this disorder obsess over matters hardly worthy of their attention; they're compelled to perform absurd acts (e.g., they're compelled to snap their fingers if they have a thought about the number seven); they're also compelled to perform normal acts an abnormal number of times or for an abnormally long period (e.g., they're compelled to brush their teeth 20 times a day or to shower for hours on end). But sufferers of OCD *know* that their obsessions and compulsions are irrational. They succumb to the compulsions *not* because they believe there to be any objective reason to do so, but because, for reasons unknown to them, they suffer intolerable anxiety if they don't. And they know that their obsessions are not legitimate; they know that those obsessions aren't comparable to Einstein's obsession with identifying framework-invariant physical laws or Beethoven's obsessive dedication to perfecting the Hammerklavier Sonata. They obsess despite themselves. They don't think their obsessions have any legitimate basis. They just can't stop them.

Thus, those afflicted with OCD don't identify with their symptoms. Unlike schizophrenics, they see their symptoms *as* symptoms. Their symptoms belong to *their minds* but not to *them;* they are like alien-invaders that have taken root in their psyches.

Bearing these points in mind, consider the following scenario. You hate

numbers. You are not detail-oriented. You want to spend all day writing novels. (You're a very talented novelist.) You know all of this. But one day you are overcome with an intense desire to become an accountant. This desire came out of nowhere. It has no roots in your personality.

You obviously don't identify with this desire. This is because it has no roots in who you are. And the reason it has no such roots is that your existing values, desires, aspirations, and so on, had no hand in giving rise to it.

Were you to act on this desire, there is a sense in which you'd be acting freely, but there's a deeper sense in which you'd be acting unfreely. You'd be acting (un)freely in a sense similar to (though not quite coincident with, as we'll see) the sense in which somebody who hands his money over to a gunman is acting (un)freely.

Suppose that, while pointing a gun at Smith, Green says "your money or your life." In handing over his money, Smith is acting freely in the sense that he's doing what it is that, the circumstances being what they are, he wants to do; and he's acting unfreely in the sense that, the circumstances being what they are, what he wants to do is not what he wants to want to do.

Much like Smith, you're damned if you do and damned if you don't. Supposing that you knuckle under to your ego-dystonic desire to become an accountant, you are acting freely in the sense that you are doing what you want to do, and you are acting unfreely in the sense that you are not doing what you want to want to do.

That said, there is a sense in which you are less free than Smith. The circumstances being what they are, Smith must exercise his agency in ways that he'd rather not have to exercise it. But this agency is quite intact.

But *your* agency is not intact. Your ego-dystonic desire to become an accountant is competing for control over the very mechanisms that convert ego-*syntonic* intentions of yours into action; and were you to give into this desire, you would to that extent be forfeiting control over those mechanisms. By virtue of having that desire, you're in a position similar to that of the President of a country much of which is controlled by war-lords who don't recognize his authority—warlords who, though they operate *in* his country, are not *of* it, and whose sphere of authority is only as big as the President's sphere of influence is small.

Smith's situation is like that of a President whose country is being attacked by a foreign nation but who has complete control over his own country. Even though Smith is in a situation in which he is forced to act in

ways in which he wouldn’t otherwise act, his autonomy is not on that account compromised; and there is thus an important sense in which Smith is freer than you are.

4.5.1 A point concerning mental illness

Earlier, I referenced Beethoven and Einstein. Interestingly, both suffered from OCD.[456] So do many other high achievers. What this shows is not that OCD *per se* conduces to high performance, but that the personality traits that, when pathologized, are expressed as OCD lend themselves to high achievement. Beethoven, Einstein, and other OCD-afflicted high achievers were capable of intense concentration, were highly self-critical and, in order to satisfy the draconian standards of perfection they imposed upon themselves, were punctilious in their prosecution of the projects they undertook. Someone with these performance-enhancing character traits is more likely than others to respond to anxiety and disturbances generally by becoming obsessive and compulsive. But *mere* obsessiveness and compulsiveness are not pathological.

Also, there isn't a well-defined line between pathological and non-pathological obsessive compulsiveness; and one's pathologies may be implicated in, and in some cases even enhance, one's efforts. So given that Beethoven had OCD, it's unlikely that it was to no degree involved in his composing the Hammerklavier Sonata, even though his composing that work was, on balance, non-pathological.

This exposes one of the main differences between physical and mental illness. Physical illnesses are to no degree adaptive (except in some purely accidental manner—one's having asthma might prevent one from joining the army and then being injured on the battlefield[457]). But mental illnesses, specifically ego-dystonic ones, are sometimes adaptive. This is because ego-dystonic illnesses are what result when defensive mechanisms become too strong. Excessive insecurity, fearfulness, and anxiety are defense mechanisms that have gone too far and that, consequently, inhibit the patient in some respects. But those same overdeveloped mechanisms may enhance the patient in other respects. An OCD mathematician's ability to focus on a minute task for hours on end is conducive to real achievement. In that respect, it's non-pathological. And there's no real chance that the psychological mechanisms that make it possible for the mathematician to sustain such a high level of concentration for such a long time have a *purely* pathological basis. But it's a distinct possibility that, depending on the mathematician, the mechanisms that make that possible are considerably reinforced by mechanisms the very purpose of which, at least when they were first developed, was to narrow the

individual's psyche, so as to cast intolerable conceits into the abyss of the unconscious.

It should be pointed out that mechanisms whose *original purpose* is neurotic can end up becoming non-neurotic virtues of character. For example, to deal with some specific problem (e.g., a sense of helplessness that is induced by an unstable home-life), a given person might over-develop some existing character trait (e.g., a tendency to focus on details becomes a maniacal obsession with them), which, after the initial disturbance has been dealt with, assumes some non-pathological, adaptive form (e.g., a talent for neurosurgery).

Basically, there isn't always a clear distinction between mental illness and mental hyper-wellness. And the fact that the border is blurry isn't to be understood in terms of the trivial, ubiquitous fact that most borders are blurry (that, for example, there is no sharp line between short and tall, poor and rich, etc.). The blurring of the line, where mental illness is concerned, has to do with the fact that such illness is, in some ways, at certain junctures, constitutive of certain kinds of mental health (or, what may be different, mental flourishing).

This, I think, is why "genius" and "madness" are proverbially thought to go together. I *don't* think that the two, in fact, go together. Madness is always performance-inhibiting, for reasons that we'll soon discuss. But the eccentricities associated with ego-dystonic ailments are often conflated with madness, leading to the false view that insanity is associated in some way with genius. I must emphasize that ego-dystonic ailments are *not* a form of madness. (Not all psychological disturbances are forms of madness.) So far as the word "madness" has a clear meaning, it refers to a condition in which one *identifies* with one's pathologies—it refers to ego-syntonic ailments. A precondition for such conditions is a lowering of inhibitions that, in sane people, remain operative and that, when *hyper*-operative, constitute OCD and other ego-dystonic ailments. Goblins and ghouls are no part of a sane person's life except when, during sleep, there is a lowering of the inhibitions that, during wakefulness, rightly keep these creatures of childhood in the cellar of the unconscious. A person with OCD (who isn't also a person with an ego-syntonic disorder) is like the sane person, except that the OCD person's defenses are in overdrive. So, in such a person, it isn't just ghouls and goblins that get the ax: healthy beliefs and sentiments do as well. One's antipathy towards one's father and one's suspicions concerning one's best

friend, become off-limits in the way that, in a less neurotic person, only goblins and dragons are.

In terms of its outward effects, this hyper-inhibitedness can mimic "madness" (which term, so far as it means anything, refers to psychosis). The psychotic person refuses to touch the doorknob because he thinks that it's Satan's horn. The obsessive-compulsive person refuses to touch it because, in his mind, it's connected, by way of emotionally charged associations *that he knows to have no objective basis,* with emotions that he doesn't countenance (e.g., urges that, at some level, one believes to be are wrong, such as incestuous or sadistic urges towards a loved one). But the outward observer only sees some eccentric who refuses to touch a doorknob. Hence the conflation of psychosis with ego-dystonic illness; and—since the latter is associated with (though it's not the most productive manifestation of) character traits that are associated with high achievement—hence the bogus stereotype of the "mad genius."

Ego-*syn*tonic illnesses, unlike ego-dystonic ones, are not at all adaptive, the reason being that they result from a *weakening* of one's defenses to a profound, highly incapacitating disorganization of one's psyche. So, in a way, those suffering from ego-dystonic illnesses are even further removed from those suffering from ego-syntonic illnesses than those who suffer from neither. In this connection, it is worth pointing out ego-dystonic maladies seem to coincide with what are sometimes called "neuroses" and that ego-syntonic ones coincide with "psychoses."

Chapter 26 Selves vs. Minds, Values vs. Desires.

1.0 An as of yet unidentified problem with Frankfurt's analysis

According to Frankfurt, a free act is one that is driven by higher-order desires. This is false. A free act is one that is driven by one's *values*. Values are not desires. (Values may lead to desires, and desires may lead to values. But values aren't desires.) A story will help us begin to establish these claims.

You hate opera, and you hate the kind of people who hate opera. But those people are chic; they're the "right" sort of people. You want to be one of them; you want to be chic. But it isn't because you see their lives as being valuable that you want to be one of them. In fact, you see them as empty, bad people who act in empty, bad ways. But you're a follower, not a leader. Your desire to enjoy listening to opera has nothing to do with your valuing opera or, indeed, with your valuing anything. That second-order desire of yours is simply a means to end, the end being the fulfillment of some first-order desire, the nature of which we will discuss in a moment. You know that, if you don't come to have a genuine love of opera, you'll inadvertently disclose to the people whose ranks you are trying to join that you don't much care for them. While in your box-seat at the opera-house, you'll look bored and sullen; you'll sound fake when you tell these big people how much you like opera. And you believe that, as a result, they won't let you join them, in consequence of which you'll wither away.

That first-order desire is not a desire to survive: you know that your survival doesn't depend on your being in the good graces of these people. Rather, it is a cowardly desire to fit in. And *that* first-order desire is, we may plausibly suppose, the off-spring of another, equally despicable first-order desire, namely, your desire to avoid the hard, anxiety-ridden work of forging your own identity. So it is laziness and cowardice that underlie your first-order desire to fit in. It is not, I repeat, a desire to survive.

In this context, your behavior is a compulsive response to fear and laziness. It is not a free act. It is a compulsive, pseudo-free act.

In this context, it is important to distinguish between fear-driven behavior and cowardly behavior. People are vulnerable. They must take their vulnerabilities into account when deciding what to do. The coward has no compunctions about parting with who he is in his efforts to survive. The fearful, non-coward will avoid parting with his identity *until*, having considered all of the possibilities, even those that would involve the hard work and alienation from ways of living to which he is accustomed, he judges that he has no alternative but to forfeit at least part of his identity in order to survive. It should

be said that many a fearful, non-coward will risk death before forfeiting his identity or even a part thereof.

1.1 Second order desires sometimes are the puppets of first order desires

So, in this context, your higher-order-driven desire is purely instrumental: It is just a means by which you try to gratify your desire to survive. And, though you *want* to survive, you don't *value* it; or, what is probably more accurate, your desire to survive is more basic than your judgment that it would be good to survive, and the latter is an intellectualization, a derivative, of the former. Your desire to survive is a force unto itself. It doesn't await your *approval*. It directly motivates action. And, should you come to approve of it, this judgment of yours has little role, if any, in the instigation of actions on your part that, so you hope, conduce to your survival. Your desire to survive is much more primitive than any judgment of yours as to the extent to which some norm is being complied with. All animals want to survive; but not all animals make judgments. Even though there may well be cases where your desire for such and such is a consequence of your judgment that such and such has value, your desire to survive isn't one of these cases. Thus, second-order desires can be value-free; for they can be nothing more than the techniques for gratifying (valuatively empty) first-order desires.

1.3 Fake freedom: The warping of one's higher-order desires to make them validate one's lower-order desires

What if one *warps* one's higher order desires to validate one's lower order desires? What if one rationalizes? When lecturing on Frankfurt's analysis of freedom, a student of mine, Jared Murphy, asked these questions. Until that time, I had accepted Frankfurt's analysis. But Mr. Murphy's question made me rethink it and then reject and replace it.

These questions show that, as it stands, Frankfurt's analysis won't do. Knowing that doing right by one's existing second-order desires would be difficult, and knowing that doing right by one's first-order desires would be easy and pleasurable, one can convince oneself that one's current second-order desires are the wrong ones and that they ought to be replaced with ones that, as it happens, validate one's current first-order desires. If one is too lazy and weak to do what one believes to be the right thing, one has a choice: (i) one can choose to do the right thing, which, as one knows, involves making much needed, but labor-intensive, changes to one's own character; or, (ii) one can choose to replace one's existing second order desires with ones that---what a surprise!---just happen to validate one's existing first-order desires.

Doing this involves convincing oneself that one's character-defects are actually virtues. So, supposing that Green is somebody who is doing just this, here is what he might tell himself ("GR" is short for "Green's rationalization"):

> (GR) I'm not "lazy." Unlike the blinkered maniac who works all the time and never takes the time to smell a rose or go to the beach, I see the larger picture. Nor am I "weak." Unlike the so-called "hero" who moronically tries to buck the system, I see the larger picture; I see that despite that system's imperfections, it is one that I am duty-bound to comply with; and, unlike the so-called hero, I am enough of a realist to know that complying with the system involves getting one's hands dirty. The "hero" isn't a hero at all.
>
> To be sure, the (so-called) hero's efforts have a sheen of valor. And, in sticking up for the little guy---the one whose work is being plagiarized by the well-entrenched senior professor, the one who is being framed by the oft-decorated police-officer---the hero's conduct might initially appear to be more in keeping with

the rules that are constitutive of the system. But that initial appearance is an illusion.

We need the system. If we don't follow the rules, the system will collapse. But there are two kinds of system-constitutive rules: those that are spoken and those that are not. The hero's behavior is heroic only relative to the first set of rules. But, in this context—the one in which it would be said by a short-sighted person that I am being "weak," the one in which I'm supporting the senior professor's portrayal of events---it is the second set of rules that count. The hero (the odd graduate student or pre-tenure professor who is "telling it like it is" and isn't abetting the efforts of the senior professor) is violating those rules. And he's doing so *knowing* that, because the rules he is violating must never be spoken, people will be forced to characterize his anti-social, system-undermining behavior as system-compliant. So the hero's behavior is particularly insidious. Therefore, to the extent that his behavior is antithetical to mine, I will take it as a complement when somebody reacts my system-friendly conduct by calling me "a weak, faceless bureaucrat" or "an empty, servile, cringing shell."

To be sure, the system isn't perfect. It often rewards the wicked and punishes the virtuous. It often requires that one speak ill of somebody who one knows is decent---who, though nobody says it, everybody (except for a few imbeciles) knows is decent and who, though nobody says it, everyone (same qualification) knows to be a person of merit---and to laud somebody who one knows (and who everyone knows, etc.) to have no merit.

But, on balance, the system is good, and to keep it going, one must follow the rules---*all* of them, not just the spoken ones. And if that means taking the "coward's path," so be it.

Deep down, Green doesn't really believe GR. Deep down, he knows that he's a servile, weak person who is doing the pragmatic thing, as opposed to the right thing. But, at least to the extent that he's convinced by GR, Green has replaced one set of higher-order desires (e.g., a desire to act virtuously, even when, because of circumstances, one wishes to act in a craven and iniquitous manner) with a different second-order desire (a desire to resist the odd

temptation one occasionally feels to do the “right” thing, even when it is not to one’s practical advantage to do so).

It’s pretty clear that this world is replete with people like Green.[458] This is a problem for Frankfurt’s analysis. So far as Green is acting on his *new* higher-order desires---the ones that, in telling himself GR, he is installing---Green is letting his own cowardice make his decisions for him and he is thus *not* acting freely. A junky who lets his low cravings make his choices for him; he is acting freely only when he takes the reins and acts in accordance with what *he* wants to do, as opposed to what his *addiction* wants him to do. Green is similarly unfree, since, instead of doing what *he* wants to do, he does what his *cowardice* wants him to do.

So, contrary to what Frankfurt says, an act can be *unfree* even if it expresses a higher-order desire. This happens when the higher-order desire is the puppet of a lower-order desire.

2.0 Values ≠ desires

Given the points just made, Frankfurt's analysis of freedom must be modified or augmented in some way. More specifically, Frankfurt must take one the following four positions (and make any such adjustments to his system as the position chosen requires). The result will, I believe, be a decent analysis of at least one, very important kind of freedom. Here are those positions:

> (i) Contrary to what was just said, people *don't* rationalize; they *don't* deceive themselves.
>
> (ii) Although people do rationalize and deceive themselves, their doing so doesn't involve their *actually* replacing any higher-order values of theirs with other higher-order values. It involves only their falsely *believing* that they have done so. The original values stay on, this being why rationalizers are tormented by gnawing feelings of guilt (from which they typically attempt to insulate themselves by adding more rationalizations to their existing rationalizations). In other words, rationalization isn't about *actually* replacing one's values/higher-order desires with new ones. It is about making oneself *believe* that one has effected such a replacement.
>
> So, in deferring to the senior professor, Green *isn't* acting according to his second order desires. Rather, he's acting according to what he's deluded himself into believing are his second order desires. (Consequently, the fact that Green's actions are not free, except in the sense that they're voluntary, is consistent with Frankfurt's analysis.)
>
> (iii) Situations such as those described by GR *neither* constitute cases where one set of higher-order values is replaced with another *nor* constitute cases where some new set of higher-order values has altogether failed to have been installed. Rather, such situations involve a bifurcation or fracturing of the psyche---a kind of splitting of oneself into two selves, one of which still has the old set of values and the other of which has the new set of values.

So, according to this hypothesis, rationalizations *do* install new higher-order desires. Contrary to (iii), however, they do so, not by ousting old higher-order desires, but by adding new ones to them.

Given that the new ones are inconsistent with the old ones, rationalization necessarily involves a degradation of one's psychological integrity. One's thinking must become more illogical than it used to be *or* one's agency---by which I mean the totality of faculties involved in one's being able to decide what to do---must undergo a sort of splitting.

(iv) Although second-order values can be uninstalled and replaced with others, a distinction must be made between higher-order values, on the one hand, and *core* values, on the other. All core values are higher-order values, but not all higher-order values are core values. Green's actions are unfree because, although they're consistent with (some of) his higher-order values, they aren't consistent with his core values.

2.1 Free acts = value-driven acts

(i) is out of the question. To say that people don't rationalize is itself to rationalize, given what an obvious empirical fact it is that people rationalize.

Also, the essence of Frankfurt's position is that, if one is to be free, one must not be at war with oneself: intrapsychic conflict is the enemy of freedom. Frankfurt is aware that there are different ways of eliminating intrapsychic conflict and that some of them not only fail to bring one freedom but strip one of it. For example, given somebody who is torn between shooting heroin and kicking his addiction, that person can cease to be at war with himself by relinquishing his desire to kick his addiction and, thus, to surrender to it. But, as Frankfurt himself clearly states, by resolving his internal conflict in *that* way, the junky is forfeiting his freedom. Nonetheless, Frankfurt holds that, in cases where internal conflict is so intense that no one party has uncontested control over the reigns of agency, one cannot possibly become free without ceasing to be internally conflicted. Knowing that some ways of resolving internal conflicts fail to make one free, Frankfurt holds that the way in which a person resolves a case of internal conflict will bring that person freedom if and only if two conditions are met, to wit:

> (a) The resolution of the conflict involves his higher-order desires prevailing over one's lower desires;
>
> (b) The way in which that person's higher-order desires prevail over his lower-order desires doesn't involve the latter being unduly frustrated and, consequently, requires that the victorious higher-order desires so align with the lower-order desires over which they've triumphed that the latter can discharge themselves in a way that doesn't threaten the hegemony of their higher-order masters.

This position is a near duplicate of a position that Freud explicitly advocated, in various forms, throughout his career.[459] It may be possible in some attenuated, strictly logical sense for there to be Intrapsychic conflict without self-deception or rationalization. But, human psychology being what it is, it seems that self-deception is essential to the perpetuation, if not the initiation, of such conflict. In any case, (i) is *prima facie* discrepant with empirical facts; and, although many philosophers and psychologists virulently deny (i), they do so on

a priori grounds that, as we'll see, are made of straw *or*, what is much more common, they don't bother to defend their position.

(ii) is inconsistent with empirical facts. People's values do change. As a 20 year old, Jim wants to save the world. He works for the Peace Corps. He spends his weekends working at soup kitchens. As a 40 year old, Jim is an investment banker who lives to accumulate wealth and wouldn't dream of not having a live-in maid, butler, etc. One *could* say that, even as a 20 year old, Jim's values were *really* those of a greedy banker *or* that, as a 40 year old, his values are really those of a hippie do-gooder. But such a position would be *ad hoc* and counter-empirical.

To be sure, there are many cases where somebody who is *in fact* a greedy sociopath works at a soup-kitchen (etc.), so as to bolster his resume in the hopes of making a lot of money on Wall St. And there are people who make a lot of money on Wall St. who wish that they were doing good for the world and whose consequent grief all the gin in the world cannot drown out. But there are also people whose values truly change.

(iii) describes something that, I suspect, really does happen. But, I also suspect, it does not describe what is *typically* happening in cases of rationalization. In my personal experience, and in the clinical experience of many an authority, what *typically* happens is that some *one* set of higher-order desires becomes corrupt: there is a mixing of the old (pure) higher-order desires with new (corrupt, rationalization-driven) higher-order desires. There is still just one agent. There aren't two agents—a Dr. Jekyll and a Mr. Hyde—inhabiting one body. There is but one agent. Some of his higher-order desires are wholesome and some of them are unwholesome; and in some cases a single higher-order desire may be wholesome in some respects and unwholesome in others.

(iv) is the only remaining option. In the pages to come, we'll find much reason to accept (iv) and no reason to reject it. Incidentally, this analysis, though clearly similar to Frankfurt's analysis, does differ from it in a non-trivial way. Frankfurt seems to hold that one's higher-order desires *ipso facto* are one's core desires. But that is precisely what the present hypothesis denies.

2.2 Some general facts relating to the nature of values.

In the next section, we'll answer the question: "what are values?" Having answered this question, we'll be able to answer the question: "what is it for an act to be free?" We'll answer the latter question in the next section.

(1) Values must be distinguished from things that have value. A great novel has value, but it is not itself a value.

(2) "Value" is a relational term. What has value for Smith may not have value for Jones.

(3) (2) doesn't entail that no value is better than any other value; nor does it entail that values are in any way 'subjective' or 'unreal.' The term 'three miles from' is relational; but whether Smith is three miles away from Jones has nothing to do with anyone's feelings or opinions about anything. The same thing is true of values, as we'll see. Given that values are judgments, values can be correct or incorrect.

Given that values are relational judgments, two people who have different values needn't disagree about anything. I value playing the piano and I don't value playing scrabble. That means that, in my judgment, *I* bear a certain relation to the first activity that I don't bear to the second. Smith, let us suppose, values playing scrabble more than he values playing the piano; this means that, in his judgment, *he* bears a certain relation to the one activity that *he* does not bear to the other. Those two judgments are perfectly compatible. So even though my values are antithetical to Smith's, there is nothing that we disagree about.

(4) Liking something is not the same thing as valuing it. People value activities (e.g., helping somebody in need) that they don't enjoy and enjoy activities (e.g., doing cocaine) that they don't value.

One tends to enjoy the activities that one values. I value playing the piano, and I also enjoy it; and these facts are

obviously closely linked. But the relation I bear to that activity by virtue of enjoying it is distinct from the relation that I bear to it by virtue of valuing it. By virtue of the fact that I value playing the piano, I bear a *cognitive* relation to that activity; in other words, my valuing it consists in my having made a *judgment* about it (viz., I see it as having value and therefore---for reasons that we'll soon identify---as being agential). By virtue of the fact that I enjoy playing the piano, I bear a purely *affective* relation to it (viz., it makes me *feel* good).

2.3 What are values?

A value is a judgment on the part of some person (or sentient creature) to the effect that a certain course of action, or a certain outlook, will increase the scope of its *agency*. A moment ago I was reading a book (*The Scientific Image*, by Bas Van Fraassen). I then had a sudden urge to read a different book (*The Will to Power*, by Nietzsche). I was driven by a whim. Van Fraassen's book was illuminating; I was not wasting my time by reading it. And, although, for personal and professional reasons, I do at some point have to read Nietzsche's book (which, judging from the fragments I've read, is extremely good), I would, in giving in to my whim to ditch Van Fraassen, be undermining a resolution of mine, and doing so for no good reason.

The result of my ditching Van Fraassen under these circumstances would have been a degradation of psychological integrity. I would have been allowing my decision-making process to give undue weight to sudden fancies. I would have been sending a message to myself that it's ok *not* to carry out an intention that I had formed in a rational and informed manner and that I knew that, if successfully carried out, would make me a better and more powerful person. This illustrates, and also supports, though only to a small degree, our contention that value is a judgment to the effect that a certain course of action or way of living increases the scope and power of one's agency.

I value carrying out rationally formed intentions. But why? Because I would be diminishing myself if I indulged every passing whim. So far as the fickle winds of fancy determined by life-path, *I* would not be calling the shots; *I* would not be the captain of the U.S.S. JM Kuczynski. So far as impulses, rather than rationally formed intentions, determined my conduct, I would not be an *agent.*

Even though I may have *desired* to read *The Will to Power* more than I *desired* to read *The Scientific Image*, I *valued* reading the latter at that particular juncture more than I *valued* reading the former. My reason: under the circumstances, my reading *The Will to Power* would have constituted a surrender of agency to sub-agential whims.

2.3.1 Interlude: An objection to what we just said

There seem to be counterexamples to this contention. For there seem to be cases where, driven by his values, somebody martyrs himself or otherwise diminishes himself. There are cases where a person correctly characterizes an act as "valuable" even though that same act, if carried out, would diminish his agency and possibly destroy it. But that isn't because our thesis is false. It's because the word "value" is ambiguous. Sometimes it means "belief that a given way of acting (or being) enhances *one's own* agency." Sometimes it means "belief that a given way of acting (or being) enhances *agency in general*---not necessarily one's own, but net amount of agency had by people overall." This second usage of the term "value" is derivative of the first, even though one's initial impression would probably be that it's the other way around.

Most ethical systems use the word "value" (or its analogues) in the just-mentioned derivative manner. The one exception to this is *ethical egoism*, which was advocated by Nietzsche (and not---at least not to my limited knowledge----by very many other thinkers of substance). We'll find at the end of this book that ethical egoism, though not ultimately a viable moral system, has some hidden virtues.

2.3.2 What are values? (revisited)

There are times when I *desire* to watch TV more than I desire to read *Non-Standard Analysis* by Abraham Robinson. But, in general, I *value* doing the latter more than I value doing the former. (Why "in general" as opposed to "always"? There are times when my reading *Non-Standard Analysis* would represent a compulsive, pseudo-agential act; and on many of those occasions, my watching TV would be a more "integrated" and rational course of action than would my reading some onerous treatise. In general, many is the time when what would ordinarily be characterized as the "lazy" course of action is the rational, non-lazy one, the reason being that what is in fact an agency-diminishing impulse cloaks itself as an intention to perform some action that would ordinarily be seen as agency-promoting, as when one's intention to cut one's break short so as to go for a five mile run (even though one's bursitis makes it a bad idea to do so) or to continue reading some treatise (even though, owing to one's fatigue and consequent vulnerability to performance diminishing influences, one would ultimately be a better treatise-reader (and treatise–writer) were one, at this particular juncture, to take a break from treatise reading---we'll revisit this obscure parenthetical point). I *value* reading *Non-Standard Analysis* even though, at that particular juncture, I do not *desire* to do so, at least as much as I desire to watch TV, because I judge that reading that book would consolidate my agency and expand its jurisdiction, whereas watching TV would fail to have such consequences and would indeed lead to a withering of agency.

To be sure, once I turn off the TV and start reading *Non-Standard Analysis*, I not only experience gratification, but do so on a number of levels. First of all, I am proud of myself for doing the right thing. Second, and much more importantly, in reading *Non-Standard Analysis*, I experience what I will refer to as *agential enjoyment*, whereas in lazily watching TV I am experience non-agential enjoyment. (We tend to use the word "pleasure" to refer to non-agential enjoyment: one experiences "pleasure" when one enters a warm bath. But, although one *enjoys* reading difficult treatises or working on one's backhand, and although one *does* experience *a kind of pleasure* in doing so, that pleasure is derivative of a certain kind of enjoyment (for lack of a better word) that comes from exercising one's agency.)

2.3.4 What are values? (continued)

I *have* impulses. Impulses are *in* me. But they are not *of* me. *I* am not my impulses. There is a sense in which, so far as my behaviors are impulse driven, *I* am not acting. To put it another way, so far as I *am* acting in such circumstances, my actions are but vehicles for sub-agential impulses. Consequently, my behavior in such circumstances is ultimately only pseudo-agential. In such cases, that is to say, my agency is involved, but only as a way of discharging impulses that are sub-agential and, in being acted on, diminish my agency.

Here it is important that we make a distinction between two kinds of impulses: those that one has *by virtue* of being an agent and those that one has before one's mind becomes agential. My impulse to read *The Will to Power* is obviously not in the same category as my impulse to eat tubs of ice-cream or to engage in certain carnal acts. The former impulse---though one that, in the previously described circumstances, ought to be suppressed---is obviously derivative of the development in me of values and aptitudes that jointly constitute agency. For that reason, were I, on a whim, to stop reading Van Fraassen's book and start reading *The Will to Power*, I would rightly not judge myself as harshly as I would if I stopped reading Van Fraassen's book and punched my annoying next door neighbor or ate a tub of ice-cream. Giving into impulses---be those impulses agency-derivative or agency-independent---is a bad thing. But giving into agency-independent impulses is worse.

"But by *never* giving into one's impulses," it must be asked, "doesn't one become a robot? Doesn't one become a brittle, neurotic square who can't live, love, or laugh?"

How that question is to be answered depends on what, in this context, it means to "give in" to an impulse. A healthy person ought indeed to discharge his base impulses and---what sometimes coincides with so acting---one ought to act spontaneously. But when one scrutinizes a non-maladaptive case of "giving into one's impulses" or of "acting on the fly," one finds that the decision to do so was a considered one; one finds that the thought-process behind the decision was to the effect that, since one damages oneself and deprives oneself of strength by alienating oneself from one's base impulses---of which one's higher values are rarified derivatives---one weakens oneself: one turns oneself into an emotional cripple, a neurotic nerd, who, having drained his agency of a source

of power, is less agential than one would otherwise be. So what *appear* to be cases of acting "on a whim" are cases in which one non-whimsically (though perhaps almost instantaneously) realizes that it is to the advantage of one's agency to act in a way that *seems* to be whimsical. *Genuine* cases of whimsical behavior are the prerogative of schizophrenics, whose minds are so disorganized that impulses don't have to go through the usual vetting process to be converted into action, and of psychopaths, whose vetting process operates much more intermittently than that of a normal person and is governed by different operating principles. (Though it may not be apparent, these points about psychopathy are in the process of being substantiated.)

2.4.0 The question “what are values?” considered in light of Mill’s analysis of moral value

According to John Stuart Mill, happiness is the one thing that truly has value. [460] So far as anything other than happiness has value, Mill says, it's only to the extent that it leads to happiness. Right now we'll examine this thesis of his, along with his argument for it. In so doing, we’ll find support for our contention that to value *x* is to judge *x* to increase the scope or the strength of one’s agency.

There’s no denying that writing novels of the same level of merit as *War and Peace* is more valuable than doing cocaine. Mill admitted this. Mill also admitted that, not having been duly educated or endowed with the requisite cognitive faculties, one might fail to know the delights of reading *Hamlet* and would prefer drinking vodka to reading Shakespeare. But, Mill insisted, so far as reading *Hamlet* is more valuable than drinking vodka, it is *only* to the extent that the former brings more happiness than the latter.

One problem with what Mill says is that happiness *presupposes* a value-system. Happiness involves believing oneself to have measured up to certain benchmarks and thus to embody certain *values.* (Mice can probably feel good. But they can't be *happy*, they can't make judgments of the relevant kind.) A consequence is that happiness *presupposes* that there are things other than happiness that are of value.

Thus, Mill's contention that happiness is the one and only non-instrumental good is false. In fact, it's tautologously false, given that happiness consists in judging oneself to be good in some way or other. What is good for one person may not be good for another.

2.4.1 The question “what are values?” considered in light of Mill’s analysis of moral value (continued)

Also, Mill claims that, if a person fails to prefer Tolstoy to grain alcohol, it's because he's defective in some way. But in making that claim, Mill is saying that Tolstoy is valuable in a way that grain alcohol is not and that one must be deficient not to see that and, therefore, not to derive more pleasure from reading good literature than from drinking cheap liquor. So Mill is saying that *there are* values other than happiness and that our happiness is a *consequence*, and thus not the essence, of our receptiveness to these values.

Of course, Mill could claim that one can fail to derive any pleasure at all from Mozart's music, etc. and yet *not* suffer from any defect. But in addition to being totally *ad hoc*---not to mention hypocritical, at least when put forth by the likes of a cultured person such as Mill, who spent his days reading “good” literature and listening to “good” music---such a position is brazenly false. A tone-deaf person who can't distinguish Mozart's music from the noise made by a dump-truck clearly is deficient in some way: there is something to which he is blind (or deaf, rather). Setting aside people who are rightly held in contempt for their musical deficiencies, nobody really believes that any given person is as good a composer as Mozart or, equivalently, that Mozart is as bad a composer as any given person. At least some of Mozart’s works are better than at least some of those of at least some other people. This is a conservative claim. It’s hard to see how one could sincerely deny it.

This is not to say that, given *any* two musical works, one is better than other. Two musical works may not be comparable. Each may be good in its own right and it would be artificial to say that one was better than the other. But there are aesthetic truths. I wrote a fugue once. It wasn’t as good as any of the fugues that Bach wrote. The statement “it’s *not* all a matter of how we feel about a given work of music: some musical works really have merits that others lack” doesn’t entail “there is some *one* relation *R* such that musical work *x* bears *R* to *y* iff *y* is better than *x* and such that, given any two musical works, one bears *R* to the other." (Aesthetic properties are not unusual in this respect. Outside of the narrow horizons within which we operate, the question ‹is *x* hotter than *y*?› ceases to have a determinate meaning or therefore a single answer. But surely an object’s thermal properties are not “subjective” or “unreal.”)

Of course, these days, many *claim* to reject this sort of non-relativism, and to

hold its advocates in contempt. But that can very easily be explained without supposing either that it's true or that anyone believes it. If it's all just a matter of taste, then—good news!---your sonatas (which even your mother describes as "execrable") are just as good as Mozart's. We're all geniuses! Break out the party hats!

If there is no aesthetic right or wrong---if a person's view as to the level of value of a given musical work is a projection of his emotional (and, so it is assumed to follow, unfounded) response to it, instead of his emotional response to it being a consequence of his belief as to its degree of merit then, at least in the arena of musical composition, nobody need fear censure. And given how obviously false that view---or, at least, how thoroughly at variance it is with people's behavior and sentiments---and given, therefore, how much self-deception is involved in making oneself believe it, that view makes it possible to accept other positions that absolve one of one's deficits. No longer is one stupid; one just has a different "but equally valid" opinion.

Mill could rid his position of its vicious circularity by replacing the term "happiness" with the term "pleasure," in which case his thesis becomes "a thing is valuable to the extent, and *only* to the extent, that it promotes pleasure." An immediate consequence of this view is that the life of a tapeworm that experiences a large amount of pleasure is more valuable than that of an Einstein who does not experience quite as much pleasure. Mill is aware of this problem with his view. He deals with it by saying that listening to Mozart, reading Shakespeare, etc. brings one more pleasure than does shooting heroin. But while it can be said that reading *Timon of Athens* brings one more *happiness* than does shooting heroin, it cannot plausibly be said that it brings one more *pleasure.*

2.5. An objection to our position concerning the nature of values

People sometimes see things that diminish them---that abridge the scope of their agency and possibly even kill them---as being "valuable." In some cases, this is simply the result of error. (I value doing push-ups all day because I foolishly think that, as long as I do so, I'm in control of my life.[461]) But in many cases no error is involved. Smith *rightly* sees his self-diminishing activities as a UNICEF-volunteer as having "value." In fact, practically every system of ethics *demands* a reduction of oneself. [462]

This objection is not hard to neutralize. Let VS be a value-system according to which one's primary ethical obligation is to augment the scope of other people's agency. One has no moral obligation to promote what has no moral value. So, one has no moral obligation to promote welfare (agency) unless agency has moral value. *A fortiori* one's primary moral obligation cannot be to promote other people's welfare unless agency is the most valuable thing. But agency is, of its very nature, something that has only a *relative* value: *relative to me*, my agency may well be the most valuable thing. (In fact, I have no doubt that it is.) But *relative to you*, my agency has no value at all--except in an irrelevant, purely instrumental sense. (All references to "value" in this context are to intrinsic, not instrumental, value.) So VS is saying that, ethically, what is worthless to a given individual is the very thing that has the most worth. And I do not see how such a value system could be a coherent one.

This suggests that one of the following propositions holds:

> (i) What is of moral value may not be of the slightest value to any given individual, in which case anyone who acts morally is *ipso facto* acting irrationally. (In fact, even in cases where what is moral happens to coincide with what is valuable to a given agent, it is not *as* somebody who is acting morally that he is acting rationally. His behavior, if rational, merely *coincides* with that of somebody who is genuinely moral. It does not itself embody any morality, being nothing more than a mere simulation of it.)
>
> (ii) What is of moral value is what it is in one's *own* interest to do.

So a certain incoherence is internal to *public* value systems. But such systems *must* be incoherent in the just-described ways *if* they are to do their job. Public value systems exist to ensure that, on balance, people's rights aren't horribly violated. The only value systems that can accomplish this are ones that, from a logical viewpoint, are incoherent. (There is nothing incoherent in the supposition that a value system must be incoherent if it is to have the causal properties that a public value system ought to have. This is because how much merit a value-system has is to be judged by the effects on people's lives through implementing it, not by its coherence. Value systems are not scientific doctrines. Thus, a welfare-maximizing viewpoint may be incoherent.) Setting aside people who are simply deluded (e.g., people who believe that joining cult *X*, a precondition for which is wire-transferring their life-savings to the Master's Swiss Bank account, is the royal road to self-actualization), when people say that performing self-diminishing acts (e.g., working at a soup-kitchen) has "value," they may be speaking the truth; for what they are saying is that performing such acts, though not *intrinsically* valuable, is valuable *in that* it consolidates value-systems that are instrumental in preserving and augmenting *agency in general.*

3.0 Minds ≠ Selves

In this section I will make some points about the difference between *minds* and *selves.* What I say in this section will be obscure, and I will provide little or no argumentation for it. The contents of this section will be both clarified and justified in the sections following it.

All selves are minds, but not all minds are selves. A self is a mind that is an *agent.* An agent is a mind that *acts*, and doesn't just *react.* In order for a mind to be capable of acting, instead of just reacting, that mind must do more than simply facilitate the gratification of primal urges; it must do more than provide the organism with information concerning the external world that makes it clear how, without being killed or injured, that organism can eat, sleep, etc. A mind whose only objectives are to gratify such urges, and in which the sole purpose of knowledge is to enable such urges to be gratified, merely *reacts*, in light of the knowledge that it has, to the directives of instinct. Such a mind does not act. It does not itself do anything. It merely reacts to instinctual pressures.

These reactions are not blind. The exact nature of these reactions is conditioned by the knowledge stored in that mind. But non-blind reactions are

not actions; and given an organism of the just-mentioned kind, that organism's *self* does nothing. It has no self, as we'll see; and its instincts, not its non-existent self, are the real agents. The behaviors in which the organism engages are compulsive responses to instinct. They are therefore not *acts*. Put another way, so far as those behaviors *are* acts, they are not acts on the organism's part, but rather on the part of its instincts.

The organism's mind doesn't act. Its only function is to determine, in light of its knowledge of the external world, what exact form the organism's instinct-directed behaviors must take if those behaviors are not to lead to the organism's being injured or killed. Such a mind's function is not to act, but only to ensure that the organism's *reactions* to its instincts don't run afoul of the forces at work in the external world.

3.1 Minds ≠ Selves (continued)

A *self* is a mind that has values; it is a self whose behaviors are driven beliefs as to what is of value. Value-driven acts often involve the gratification of instinctual urges; indeed, if value-driven acts systematically make it impossible for such urges to be gratified, the value-system underlying those acts is not a viable one. Nonetheless, a self is a mind that permits the organism housing it to behave in a given way only if, in its judgment, that behavior is in compliance with some value. Of course, its judgments as to what is of value are, and should be, heavily influenced by its knowledge of what will permit the gratification of primal urges. But what is relevant is that, in an organism that has a self, as opposed to a mere mind, it is *judgments* as to what is of value that lead to behavior; and it is not, except in the indirect sense just described, urges *per se* that do so.

It follows that selves *act*. Of course, creatures that are selves sometimes merely react. But it isn't *as* a self a creature reacts. (I am a tennis player. But it isn't *as* a tennis player that I am writing this book.)

3.2 Minds ≠ Selves (continued)

A story will help us move forward. Smith is a sapient being. (It doesn't matter whether Smith is human or not.) Smith can make judgments as to how it must act in order to discharge its instinctual energies. Like all judgments, these judgments involve knowledge of principles. So there must be principles that Smith is aware of and that guide Smith's thinking. But given an arbitrary one of these principles, Smith doesn't know that he has knowledge of that principle.

One day, however, Smith becomes aware of at least some of the principles that guide his thought. (*How* Smith could come to have this second-order knowledge is discussed in Sections 5.0-6.2 of the present chapter. Let "*P*" stand for an arbitrary one of the principles that Smith thereby comes to know himself to know. Let us assume that Smith has these realizations at time *t*. (Though they're experienced as mere feelings, embodied in these responses are rational judgments as to what to do, given what Smith's objectives are (instinct-gratification) and given what he knows about the circumstances. We must distinguish the instinctual response *per se* from the conscious experiences that merely represent it, much as the icons on your desktop must be distinguished from the patterns of electrical activity to which they correspond.) Before he knew that he had knowledge of *P*, Smith obviously couldn't make judgments of the form: my judgment that such and such is consistent with my knowledge of *P*. Prior to time *t*, Smith's judgments were indeed guided by a knowledge of *P*; but he obviously could not, as of yet, make judgments as to whether his thinking was being *properly* guided by his knowledge of *P*. (One's thinking can be *mis*directed by one's knowledge of a legitimate principle. This may happen either because one's auxiliary assumptions are wrong (e.g., knowing that gross disparities in wealth lead to injustice and are themselves unjust, but not knowing that certain ways of eliminating such disparities may lead to even greater injustices, President Jones destroys the economy in his attempt to redistribute wealth) *or* because one's ability to deploy one's knowledge of the principle in question has been blunted by garden-variety performance-inhibiting factors, such as fatigue and lack of memories.) But, as of time *t*, Smith can form judgments as to whether the manner in which he arrives at judgments is consistent with *P*. He is thus aware of a *norm* from which his thinking mustn't deviate; and, when such

deviations occur, he can make judgments as to their nature and extent. He can thus evaluate his own judgments. And his evaluation of them is positive (negative) to the extent that, to his knowledge, they (fail to) comply with *P*.

Integral to the argument just put forth is a delicate but important distinction. There is a big difference between being aware of a principle and being aware that one is aware of it. A talented athlete's behavior embodies awareness of principles that the athlete doesn't know himself to be aware of. The point being made here is not that his behavior *accords* with some principle of which, at every cognitive level, he is ignorant. It is that, although a knowledge of a given principle may be what is guiding that behavior, the athlete in question may not be aware of that fact and, indeed, is exceedingly unlikely to be.

Philosophers and psychologists often regard an act as being entirely "mechanical" if it isn't guided by knowledge that the agent can articulate. But it is only when one has second-order knowledge----knowledge as to what it is that one knows---that one can articulate what one knows; and philosophers are therefore likely to regard behaviors that fail to embody second-order knowledge as failing to embody *any* knowledge and, therefore, as being entirely unprincipled. In fact, one needs a certain amount of verbal dexterity *in addition* to such second-order knowledge if one is to be able to articulate the grounds on which they make practically any one of the decisions that they make.

Of course, in saying "snow is white," I am not affirming my knowledge that I *know* that snow is white; I am talking about snow, not about my own mind. But, in producing that utterance, I am *expressing* the fact that I know that snow is white. And one could very easily know that snow is white without knowing that one knows it. One can know that snow is white without having the concept of knowledge. But one must have that concept to know that one knows that snow is white. So knowledge of *P* is different from, and needn't involve, knowledge that one has knowledge of *P*. Also, by virtue of having a concept of knowledge, one has various concepts---e.g., belief, truth, mental activity---not had by an otherwise comparable creature that doesn't have the concept of knowledge.

3.3 Minds ≠ Selves (continued)

Now endowed with knowledge, not just of *P*, but of the fact that he has knowledge of *P*, Smith doesn't react to situations; he reacts to his reactions to situations. Given a situation, he instinctively makes a judgment. Aware that his judgment must comply with *P*, he makes a *second* judgment. This second judgment doesn't concern the situation in question, at least not directly; it concerns his initial judgment concerning that situation. It concerns the extent to which his initial judgment is consistent with the principles his knowledge of which eventuated in it.

Assuming there to be any cognitive dimension to it at all, one's response to any given situation is *ipso facto* principle-driven. But that, by itself, isn't enough for that response to be an action, as opposed to a reaction. Actions are behaviors that result from judgments about judgments. Confronted with a situation, one makes an initial judgment. One may then make a judgment as to whether that initial judgment was consistent with the principles guiding it. In other words, having made an initial judgment, one may *evaluate* that initial judgment. One is acting, as opposed to reacting, if one's behavior expresses such an evaluation. One is reacting, as opposed to acting, if one's behavior expresses one's initial judgment.

There is no self where there is no ability to act---where there is no agency. There is no agency, where there is no self. There is no self where there is no self-evaluation.

To put this last point in a way that, in addition to being more precise, doesn't make it seem viciously circular: there is no self where there are no judgments as to the integrity of other, co-corporeal judgments.

Where there are such judgments, there are values. To approve of (criticize) a judgment (or anything else) is to judge it to have (lack) some value that it ought to have.

What *directly* leads to a *bona fide*action is never a desire and is always an attitude of *approval* towards a desire. Where there are no values, there is no approval or disapproval. So where there are no values, there is no agency.

4.0 Co-corporeal, informationally integrated mental states enough for a mind, but not for a self

Anything that has the cognitive wherewithal to have a series of mental states *ipso facto* has a mind. But there is more to having a self than there is to being capable of having a series of mental states. Consider Smith as he is *before* time *t*. Each of his mental states could be understood on its own terms: as a reaction to immediate stimuli. Each of those reactions may have been intelligent. (In other words, each resulted from cogent reasoning as to how to discharge the relevant instincts, given the available information about the environment.) And, supposing that they were intelligent, each was more likely than not to incorporate the information encoded in previously occurring, co-corporeal mental states. So, if we think of Smith's pre-*t* mind as a series of links in a chain of events (or conditions), the links in that chain are bound together not merely by virtue of being co-corporeal and not merely, therefore, by virtue of the earlier being partly causally responsible for the existence of the later ones, but by virtue of the fact that (in at least some cases) the former ones bequeathed the information encoded to them to the later ones and, therefore, that it was *as* mental entities that the earlier ones (in some cases) had a hand in determining the contents of the later ones.

4.1 Co-corporeal, informationally integrated mental states enough for a mind, but not for a self (continued)

Still, co-corporeality, even when coupled with content-transmission, isn't enough for self-hood: supposing that $M_1 \ldots M_n$ is a series of co-corporeal mental states such that, for each *i* and *j*, ($1 \leq i < j \leq n$), for $M_1 \ldots M_n$ to form a *self*, as opposed to a mere mind, it is not enough that M_i be corporeal with M_j; nor it is enough that M_i help bring about the occurrence of M_j; nor, most importantly, is it enough that M_i do this in the distinctively psychological manner just described.

A story will help make this clear. There is a group of around 20,000 people, all of whom live on the same land-mass, which we'll call "*LM*." (*LM* is big enough that nobody feels crowded but small enough that nobody feels lonely and, more importantly, interactions among its occupants are frequent.) *LM*-occupants all speak the same language, and they often exchange information with one another. They also often exchange goods with, and provide services for, their fellow *LM*-occupants.

Being reasonably intelligent and well-informed people, these people tend to have accurate beliefs; and being reasonably honest people, they tend to provide one another with accurate information. Also, most of the time, they're reasonably honest not just in their communications with their fellows, but in their transactions with them. They tend not to defraud one another.

In at least some cases, the ties that *LM* occupants have to one another are deep ones. (For the moment, we'll set aside cases where *LM*-occupants have no significant ties to their brethren.) Those ties are not purely geographical: it isn't just that the parties involved live in the same place. Nor are those ties purely causal. (It isn't just that *LM*-occupants accidentally bump into other *LM*-occupants, thereby inadvertently changing the latter's lives.) The ties in question involve:

> (i) the fact that one person intentionally transmits information that is arrived in a principled manner to some other person who makes principled inferences from that information,

and:

(ii) the fact that two people may (for prudential, moral, or even libidinal reasons) benefit each other by honoring an agreement to exchange goods or services.

One fact about *LM*'s occupants must be made explicit. There is nothing *regulating* their exchanges with one another. Should one *LM*-occupant defraud another, there would be no penalties. Given that there are no laws in *LM*, the victim would have no legal recourse and he'd have to take it upon himself to see that justice be done. An *LM*-occupant who'd been misinformed by another would be in the same situation *mutatis mutandis*. It may be that, because *LM*'s occupants are so honest with one another, they're better off for the fact that their lives aren't weighed down by the presence of the regulatory entities that would be needed for there to be any formal penalties for fraud.

But what is relevant here is that, precisely because their lives are not thus "weighed down," the occupants of *LM* do not jointly constitute a *nation*, at least not in the sense in which France constitutes a nation. There is nothing under whose jurisdiction the occupants of *LM* fall.[463]

4.2 Minds that are selves and minds that are not

Bearing these points in mind, let us turn our attention back to $M_1 \ldots M_n$, and let us make some further assumptions about that series. The information encoded in M_i ($1 \leq i \leq n$) is likely to be accurate. That information, if transmitted to M_j, ($1 \leq i < j \leq n$), is likely to be transmitted in an accurate form. Moreover, the synthesizing of that information with other bodies of information (e.g., information currently being uploaded through the senses as to the nature of environment in Smith's vicinity) is likely to consist of inferences in which correctly understood correct rules of inferences are correctly applied.

But----and this is the final, and most important, assumption that we'll make about $M_1 \ldots M_n$-----*none of the members of that series constitute judgments concerning the merits of either the information-exchanges occurring among other such members or of the manner in which the beliefs encoded therein are generated*. Those exchanges and syntheses of information are not policed. It may be that each of the exchanges and inferences in question is done impeccably. But, if so, that isn't because there is anything overseeing those operations; it is simply because, despite the absence of any supervisory agency, the mental states involved in those operations happen to satisfy the relevant desiderata. By the same token, if the ratiocinative activity that generates $M_1 \ldots M_n$ is inconsistent, even systematically so, with the very principles embodied in that activity, no *internal* mechanisms exist to ensure that the needed course-corrections occur. The spuriousness of the reasoning involved in $M_1 \ldots M_n$ may well ensure that Smith fails in his attempts to gratify his desires.

But there isn't any internal agency that *prevents* Smith from making such mistakes or, therefore, that protects him from the damage to his person that such mistakes are likely to entail. There is, in other words, nothing under whose jurisdiction the various members of $M_1 \ldots M_n$ fall. Nothing polices either the manner in which members of that series exchange or produce information or the manner in which that series generates new members. Thus, $M_1 \ldots M_n$, though constitutive of a *mind* and perhaps of something that has an enduring identity in the sense in which a rock does, doesn't represent anything that has an enduring identity in the distinctive sense in which a human being does.

5.0 When minds become selves

Let's revisit, and duly extend, what we said earlier about Smith/$M_1 \ldots M_n$. At time *t*, it will be recalled, that series generates a mental state that embodies a judgment as to the fitness of (some of) the activities by which that series has thus far arrived at judgments. An internal regulatory agency is thereby born.

That agency will inhibit the conversion of impulses into actions. This does not mean that it will simply thwart those impulses. It means that if, once that agency scrutinizes a given impulse, that agency judges it to be one worth acting upon, then it will lift the inhibition. But, for the reasons given in Section 3.3., the aforementioned impulse isn't the proximal cause of the body-movements that result from the lifting of such an inhibition.

Even before $M_1 \ldots M_n$ became to any degree self-policing, the activity constituting $M_1 \ldots M_n$ embodied knowledge, not of some one principle, but of many. So for that activity to become *systematically* self-policing, it must be guided by a knowledge of many different second-order principles of the sort previously described. But it's a distinct possibility that not all of those second-order principles will be consistent with one another. As a result, an analogue of the process that led to Smith's acquiring knowledge of those second-order principles will lead him to acquire knowledge of third order principles, the purpose of which knowledge being to ensure that the totality of second–order principles guiding Smith's mentation be a coherent one---that those principles not be such that, when complied with, the result is that it is judged *de rigueur* to lift the inhibition on some impulse and also *de rigueur* not to lift that inhibition. Second-order principles that don't fit in with the rest are weeded out; they are no longer countenanced. For the same reason, second-order principles *are* countenanced that were not previously so. So far as this enterprise is successful, the result is an integrated body of principles, any two of which are consistent with each other, under whose jurisdiction falls all of the activity constitutive of Smith's decisions as to how to act. So, given the reasonable presumption that, at this stage, none of Smith's mental activity is purely theoretical, all of it being oriented towards instinct-gratification, this means that this enterprise, so far as it's successful, yields a single, cohesive regulatory agency under whose jurisdiction *all* of Smith's mental activity falls.

5.1 Summary

Just as group of people who live in the same place don’t become a state until their activities come to fall under the jurisdiction of some regulatory agency, so a group of co-corporeal, informationally integrated thoughts don’t necessarily constitute a self until *they* all fall under the jurisdiction of a regulatory agency. Before that happens, there are no actions, only reactions; after it happens, there are actions. We finally have an *agent* and, therefore, a *self*.

6.0 The concept of intrinsic value in relation to the concept of selfhood

With the inception of selfhood, values replace desires. Under the growing power of a supervisory structure of the just-discussed kind, the various constituents of that mind become increasingly unified: the structure they form becomes more and more of a self. (Incidentally, a consequence of this last statement, and therefore of our analysis, is that there are degrees of self-hood. This redounds to the credit of our analysis, since selfhood *does* come in degrees. Though clearly on independent grounds, a justification for it is given later.) Certainly, the *initial* purpose of this supervisory superstructure was to maximize the amount of instinctual gratification enjoyed by the mind falling within its power. And, surely, that always remains *one* of its purposes. Indeed, it arguably always remains its *primary* purpose (though, for reasons I'm about to state, I personally don't think this).

But that superstructure has needs of its own; and, as that superstructure develops, it becomes increasingly difficult to find an instinctual basis for those needs and increasingly necessary to understand them on their own terms.

6.1 Some preliminaries

Before we can see what exactly this claim amounts to, or, therefore, why it's true, we must make preliminary points of a general nature as to the nature of the supervisory agency in which we have claimed the essence of selfhood to lie. There are some impulses that, if acted on, would weaken the integrity of this supervisory agency. Tautologously, this agency, so far as it's in good working order, will inhibit such impulses. But little harm will come to that agency if it's *too* good at its job---if it inhibits impulses that it could, without in any way jeopardizing its own authority, allow to be expressed in action. If that agency is too good at its job---that is, if it's too conservative, if it inhibits impulses that, erring on the side of caution, it needn't inhibit—the result will be merely a forfeiture of pleasure; it will not, at least not in the short term (or even the medium term), do any damage to itself or to the structures responsible for the mentation falling in its jurisdiction.

As a result, this supervisory agency comes to be overly restrictive: it disapproves of the disinhibition of impulses that ought to be disinhibited. So far as those inhibited impulses are permitted gratification, it is only on the condition that they assume such forms that, if acted on, consolidate this agency's power over the course of mental activity falling into its jurisdiction. In some cases, it is possible for those impulses to assume such forms. (This is what Freud referred to as "sublimation." My aggressive instincts are discharged by my writing treatises that aggressively critique the views of my colleagues.) In other cases, this is not possible. This leads to what Freud calls "neurosis": an inhibition of impulses that cannot be discharged in ways that are in keeping with one's values but that, because they are so strong, must come out in some way or other and are forced to be expressed as *symptoms*. Since symptoms are unintentional, neurosis involves an abridgment of agency. One would have been better off yielding to the impulse and discharging it agentially, instead of forcing to find non-agential ways of discharging itself. (Freud makes this point in his (1926) book *Symptoms, Inhibitions, and Anxiety.* As he puts it: somebody riding a horse must sometimes let the horse go where it wants to go, lest the horse cease to let the rider have any control over it.)

6.2 Two Illustrations of Our Thesis that what is Valuable is what consolidates one's Agency

Illustration #1: We see Bach's music as being more valuable than Eminem's. (Eminem is a pop-star.) Why? First of all, the enjoyment of Eminem's music is easily traced back, by way of multiple pathways, to enjoyment of the sexual act, it being largely for that reason that we enjoy Eminem's music. By contrast, it is not because of any obvious connection that Bach's music has to the sexual act that we enjoy it. The pleasure Bach's music gives us is more akin to the pleasure given us by a philosophical or mathematical insight than it is to the (sub-agential) pleasures given to us by the sexual act. Implicated in the experience of listening to Bach's music, and largely constitutive of the pleasure it brings us, are ratiocinative abilities that have no obvious connection to sexual gratification, or to any other form of instinctual gratification, and that *do* have obvious connections to the rational operations on which the supervisory agency in question depends to consolidate and extend the hegemony it has over the course of mental activity. Bach's music is studied in conservatories; it is performed in places (e.g., concert-halls) where the operative norms of conduct require tremendous self-restraint---where one has to sit still for hours on end, where one must speak to others in a very controlled manner, and where one must display all of the asceticism involved in the transformation of instinct-driven proto-humans into instinct-suppressing actual humans.

When people say that Eminem's music is as highly "valued" as Bach's, so far as what they are saying is true, they are saying in a distorted way that Eminem's music is *liked* as much as Bach's. And there's no doubt that (at this juncture in history) there are more Eminem fans than there are Bach fans (that said, pop-stars are seldom liked *solely* for the merits of their music); and it may be that, during its brief recess from oblivion, Eminem's music is a source of more pleasure (not pleasure *per capita*, but aggregate pleasure) than Bach's music. But, in saying that Eminem's music is more highly *valued* than Bach's, one is confusing the concept of pleasure, and the related concept of desire, with the very different concept of value.

Illustration #2: People enjoy reading (and writing) philosophical treatises. To become capable of this sort of enjoyment, one must alter one's psychological structure in non-trivial ways. In order to pull off the remodeling of one's intellect needed to become at all adept at this discipline,

or therefore to derive any enjoyment from it, one must suspend gratification of urges that one would otherwise gratify, and one must do so without having any firm assurance that, in due course, one will be able to enjoy gratifications as intense as those that one has foregone. Because it takes a long time to acquire the philosophical acumen needed to make the reading, let alone the writing, of philosophical works more pleasurable than painful, and because during that long period otherwise fecund sources of instinctual gratification are off-limits, one must, if one has chosen to acquire that degree of acumen, develop a certain asceticism. One must come to de*value* (some forms of) instinctual gratification and to hyper-value the self-control that comes with becoming capable of suspending instinctual gratification.

This is not to say that the enjoyment of philosophical discourse has *no* instinctual foundation. Like Nietzsche, I believe that the enjoyment one derives from the exercise of one's intellect is a form, or at least a derivative, of the instinctual pleasure one receives from having power over others and acting aggressively towards them. But, supposing this correct, the way in which, by virtue of engaging in philosophical discourse, one discharges these primitive impulses is contingent on one's exercising faculties that, being purely ratiocinative, are much less directly connected with instinct-gratification than they are with the inhibitions of instinct so essential to the existence, and to the consolidation, of the supervisory agency in which the essence of selfhood lies.

The same point *mutatis mutandis* holds of the pleasure one derives from very sophisticated forms of music. Bach's music is passionate, and the passion embodied in it arguably has *a* sexual basis. But Bach's music expresses that passion in an extremely controlled and ratiocination-heavy manner; and the faculty underlying the ratiocinative component of Bach's music is one and the same with the faculty underlying the instinct-inhibitory ratiocinations constitutive of the regulatory agency that is one's self.

6.3 Selfhood a matter of degree

Some minds are more intact than others. In this respect, minds are like nations. The official government of Pakistan has uncontested control over some areas and only little or no control over others. Thus, the landmass coincident with the nation of Pakistan is *less* unified than it would be if the loyalties of its occupants weren't fragmented by the para-governmental agencies that occupy it. By contrast, the landmass coincident with the continental United States *can* be said to fall under the more or less uncontested jurisdiction of some one government, the same being true of the discontinuous landmass that is formed when Hawaii and Alaska are added. Therefore, the landmass coincident with the U.S. is *more* of a nation than the landmass coincident with Pakistan. Thus, nationhood comes in degrees. *x* is more of a nation than *y* if *x*'s government does a better job than *y*'s government ofunifying the occupants of the landmass falling in its jurisdiction.

It is for similar reasons that one mind can be more of a self than another. If the regulatory agency governing mind *x* has more scope and power than its counterpart in mind *y*, then *x* is more of a self than *y*.

That our analysis has the consequences that self-hood isn't equally distributed among minds is to its credit. People's selves disintegrate. And there are degrees of integratedness (i.e., of non-*dis*integratedness), as everyday experience makes obvious and as clinical observation of the mentally ill makes even more obvious.

Chapter 27 How Is Rationalization Possible?

1.0 Two kinds of repression

People routinely deceive themselves. It would be hard to exaggerate how often people deceive themselves or how discrepant with the truth are the lies that self-deceivers tell themselves. But some hold that the concept of self-deception is an incoherent one, and therefore that people never deceive themselves about anything. This viewpoint, in addition to being a case of self-deception itself, is demonstrably false, as we will now see.

But how exactly is self-deception possible? And what exactly does it involve?[464] The concepts of self-deception and repression are easily made to seem incoherent:

> (OR) [465] To repress an idea---that is, to drive a conscious idea into the unconscious---you have to know what it is that you're repressing and you thus have to be conscious of it; but if you know what it is that you're repressing, you're conscious of it and it therefore isn't repressed. Moreover, to *keep* an idea repressed---that is, to block that idea's attempts to re-enter consciousness---you have to be conscious of it; for if you are not conscious of it, it will slip back into consciousness unnoticed. But if you are conscious of it, it isn't repressed. So the concepts of repression, and therefore self-deception, are incoherent ones.

I will now try to show that there are two kinds of self-deception. For reasons that will become clear, this is a corollary of the fact that there are two kinds of *repression*.[466] In other words, there are two ways in which one can inhibit the entry into consciousness that one has and that, in some way and at some level, one knows that one has. One form of repression involves *inaction*. That form of repression involves a deliberate, and conscious, intention *not* to follow up on suggestive leads. One *suspects*; but doesn't *know*. And that's how one wants to keep it. So one doesn't exercise due diligence in the way of interpreting existing evidence or generating new evidence. Let's refer to this kind of repression as

“repression$_1$.”

So repression$_1$ doesn’t involve driving some already formed belief into the unconscious. It involves abstaining from doing the things that, so one strongly suspects, would generate the sort of evidence that make acceptance of that belief *de rigueur*.

The other kind of repression, which we’ll refer to as “repression$_2$," *does* involve already formed beliefs. The essence of repression$_2$ is *dissociation.* When trying to determine how to act, one tries to give due weight to each of the decision-relevant beliefs. So, in effect, each decision-relevant idea ‘gets a vote.’ To repress$_2$ an idea is to take its vote away; it is to decide to give it *no* weight.

Both repression$_1$ and repression$_2$ involve conflict between, on the one hand, forces within oneself of which one is aware and, on the other hand, forces within oneself of which one is unaware. But the natures of the unconscious forces, and also of the conscious forces, involved in the one case of the one sort of repression are different from those involved in cases of the other sort.

Both forms of repression involve the construction and operationalization of a narrative alternative to the one that, supposing the repressed idea to be correct (which it typically is), describes reality as it really is. In other words, both forms of repression involve *rationalization.* Without rationalization, it is hard, if not impossible, to repress an idea and also to keep it repressed. (That said, rationalization-construction is more heavily involved in ensuring that already repressed ideas stay repressed than it is in initially repressing them.)

Right now we’ll talk about “self-deception," since references to “repression” would resonate in a confusion-engendering way with the points just made. In due course, we’ll map what we say about self-deception onto repression$_1$ and repression$_2$.

2.0 How is repression possible?

The possibility of self-deception, and therefore of both repression$_1$ and repression$_2$, is a consequence of the following three facts.

(1) The mind is not a perfectly unified entity; it consists of various parts that communicate with each other only very imperfectly. The right hand doesn't always know what the left is doing.

(2). There are two very different ways in which one can be aware of a given thing. One can either have "knowledge by acquaintance" of that thing or one can have "knowledge by description" of it. See Chapter 10, Section 9.0, for definitions of these terms and explanations of the corresponding concepts.

(3) There are two very different senses in which a given mental content can be 'conscious.' On the one hand, it can be conscious in the sense that it is a *constituent* of consciousness. (Sense-perceptions are such constituents; so are itches, tickles, and other sensations.) On the other hand, it can be conscious in the sense that it is a mental entity *of which* one is conscious. Your belief that 1 + 1 = 2 is not a *constituent* of consciousness. It is "conscious" in the sense that you are, or can readily become, conscious *of* it. Of course, we are conscious of rocks, trees, and stars, but don't describe such things as 'conscious.' But that, I think, is because they are not mental entities and we thus have no tendency to co-categorize them with constituents of consciousness or with mental entities of any other kind. But since constituents of consciousness are the most obvious examples of mentality, and since we describe *them* as 'conscious,' we have an inclination to describe mental entities that are not themselves constituents of consciousness but that we are (non-inferentially) aware *of* as 'conscious.' We thus, without knowing it, use the word 'conscious' equivocally: sometimes it means 'mental entity that is a constituent of consciousness' and sometimes it means

‘mental entity that we are directly conscious of.’

2.1 How is repression possible? (continued)

Given (3), it immediately follows that there is a sense in which *all* mental entities that aren't constituents of consciousness are unconscious. Your knowledge that 1 + 1 = 2 is not a *constituent* of consciousness. So far as it is conscious, it is only in the sense that you can readily become conscious *of* it. And most of the time you are *not* conscious of it. But it exists, and is operative, even when you are not conscious of it. If you didn't know that 1 + 1 = 2, then the way in which you dealt with the world---the way in which you transacted with others, as well as the way you dealt with problems not having an interpersonal dimension---- would differ radically from the way in which, *even when you're not consciously thinking that 1 + 1 = 2 or* a fortiori *thinking about your knowledge that 1 + 1 = 2*, you in fact deal with the world. The way in which you reason is not only consistent with, but expressive of, a knowledge that 1 + 1 = 2, the same being true of the way in which you interact with people and also of the way in which you manage matters that don't involve others. But it is only during, at most, a tiny few of the occasions when one's thoughts and overt acts are guided by a conscious awareness of the fact that 1 + 1 = 2. It is only rarely that one consciously thinks about the fact that 1 + 1 = 2 or, therefore, that one thinks about one's belief that 1 + 1 = 2.

Bearing this in mind, suppose, if only for argument's sake, that your knowledge that 1 + 1 = 2 is identical with some physiological structure. It follows from what we've said that, even when you are not conscious of the fact 1 + 1 = 2, or therefore of your knowledge of that fact, this belief of yours is *psychologically* operative. Of course, given that it's a physiological structure, it's inevitably operative in the same, thought-free way as an involuntary muscle-spasm. But, even when not the object of consciousness, that physiological structure is also *psychologically* active; it's active, in other words, *as* a belief, and not just as a heap of molecules. Your knowledge that 1 + 1 = 2 *informs* the way in which you deal with the world even when you aren't consciously thinking that 1 + 1 = 2; and, were this belief of yours inoperative whenever it was blazing a trail across your consciousness, your way of dealing with the world would bear little resemblance to your actual way of dealing with it.

Given (2), it follows that you may not *recognize* a given mental content of yours for what it is. You can have thoughts about Charlemagne and you thus have a concept of him. But if you met him, you wouldn't recognize him; and so

long as you didn't figure out who he was, you would, without knowing it, have two distinct concepts of him (one based on personal encounters with him, the other based on statements of historians). Given anything that isn't itself a constituent of consciousness, one can fail to recognize it for what it is and, consequently, one can unwittingly have two distinct concepts of that thing. Propositional attitudes are not constituents of consciousness. So one could, without knowing it, have two distinct concepts of one of one's own propositional attitudes.

2.2 How is repression possible? (continued)

A story will clarify these points. You live next to the "Springfield Strangler": a prolific serial killer who, in the town of Springfield, has killed many children by strangling them. His identity is unknown to everyone (except, of course, the Springfield Strangler himself). You know this person as "Ted Jones." Since you watch the news, you know of the Springfield Killer. But you don't know that Ted Jones is the Springfield Strangler. You find Ted's company agreeable and you play tennis with him.

In this situation just described, there is some one individual *x* such that:

(i) you are acquainted with *x*,

(ii) you have knowledge by description of *x*,

and:

(iii) you know that, in virtue of these facts, you have two distinct conceptions of some one individual.

But this changes. For, at time *t*, you find out that Ted Jones is the Springfield Strangler. You see him being arrested. You hear the subsequent news reports. You consider various facts directly known to you about Ted Jones: his sudden nightly disappearances, which typically involved Ted's dragging large trunks to his pick-up truck; the stench emanating from his house, which he implausibly explained by saying that he had a clogged sink; etc. You thus realize that Ted Jones is indeed the Springfield Killer.

Just as one can, without knowing it, be acquainted with a given individual *while also* having descriptive knowledge of it, so one can be acquainted with a given one of one's own mental contents *while also* having descriptive knowledge of it. A corollary is that, given some mental entity *E* that belongs to one, one can *discover* that one actually has some attitude towards *E* that one did not previously know oneself to have. You can *discover* that the gnawing feeling of unpleasantness you experience whenever somebody mentions Smith is your detestation of Smith. Emotions are distinct from their manifestations in consciousness. Your love of your mother (supposing for argument's sake that

she's alive and that you do love her) is an enduring psychological structure; it is not an ephemeral event. But it *gives rise* to ephemeral conscious events (to pangs of remorse when you inadvertently hurt her feelings, to surges of joy when, after not hearing from her for three months, she calls you, etc.). But your love for her is not itself a conscious event. It is "conscious" only in the Pickwickian sense that you are conscious of it. Beliefs, like emotions, are enduring structures and are to be distinguished from the constituents of consciousness to which they may give rise. *All* propositional attitudes are enduring structures and are thus to be distinguished from any constituents of consciousness for whose existence they're responsible. And a given emotion, e.g., your hatred of Smith, can give rise to conscious events of very different kinds. It can give rise to gnawing feelings of unpleasantness (as when you hear that he's coming over), to intense feelings of satisfaction (as when you hear that he flunked his economics exam), etc. And given two qualitatively distinct conscious events, both of them expressions in consciousness of some one propositional attitude *x* and both of them, consequently, ways in which that emotion presents itself to consciousness, there isn't necessarily any reason why you should *recognize x* in both of those conscious events. To sum up, given only that *x* and *y* are both conscious derivatives of some propositional attitude *z*, it doesn't follow that any *one* part of one's mind has the information needed to recognize *z* in both *x* and *y*. For this reason, it follows that one can have two divergent attitudes towards *z*.

Given (1), therefore, it follows that one can have very different attitudes towards a given one of one's own propositional attitudes. And it also follows that, without knowing it, one can view a given one of one's own propositional attitudes as despicable *and* as laudatory, as deserving to be cultivated and expressed in action *and* as deserving to be annihilated.

2.3 Intrapsychic conflict in relation to repression

But self-deception is not mere *ignorance* of what one is feeling. In fact, self-deception seems to presuppose that, at some level, one has knowledge of the thing of which one lacks conscious knowledge. So given only that one can fail to recognize some one propositional attitude in two different conscious expressions of it, all that follows is that one is not completely transparent to oneself; it doesn't follow that one can self-deceive; it doesn't follow that there can be some sort of internal *conflict*---between what one knows and what one wants to know, between what is unconscious (because repressed) and what is conscious (because countenanced). A few additions to the three points made thus far will make it clear how self-deception is possible.

First of all, one must distinguish *having* a given propositional attitude (or, more generally, mental content) from *knowing* that one has it. Setting aside anything that has to do with repression, one can feel a certain way towards so and so but not know it. Knowing that one feels that way involves making a judgment about one's mental state.

Also, propositional attitudes never make an appearance in consciousness *in propria persona*---or, rather, they do so, but only in the indirect sense in which rocks and trees do. What is *in* consciousness; i.e., what is constitutive of consciousness, is at best some representative or derivative of a propositional attitude. On Monday, Smith feels a surge of anger towards his dad. On Tuesday, he feels another surge of anger. These surges of anger, let us suppose, are expressions of hatred. But the hatred *per se* isn't what entered his consciousness. What Smith is directly given are various indications that he hates his father. These indications are strong. (In any case, we can imagine developing the story in such a way that, supposing Smith sane, they left no room for reasonable doubt on his part as to whether he hated his father.) But he still has to "synthesize" them; he still has to make a judgment to the effect that they express some enduring attitude. In most cases, this is done so automatically, that he is unaware that any judgment has been made.

Those judgments are often quite wrong. We like the person towards whom we have constant surges of rage. We dislike the person who repeatedly makes us feel good. (He makes you feel good because he wants to disarm you. And he knows that by "pushing the right buttons"---buying you chocolates, taking you to the zoo, etc.---he'll generate feelings that, when *not* the result of this sort of chicanery, tend to express genuine affection. You'll know in your bones that

there is something very wrong about the situation. But you won't able to say what it is. Assuming that your views about others are mainstream and are thus hewed to the glib, emotionally-driven narratives in terms of which all but the most penetrating thinkers understand others, you won't have developed the ability to articulate your gut feelings or, therefore, to give them a form in which you or others are likely to give them the credence they deserve. Given that fact, along with the red-herrings that this manipulator is pouring into evidence-repository on the basis of which you make your conscious judgments, your verdict will be that he is a good person.

In any case, supposing that one has a given propositional attitude, what one is given is not that attitude *per se*, but various evidences thereof, and those evidences may be dispersed over time and may occur in multifarious contexts. When the propositional attitude is an emotionally sterile one---when it concerns a belief about number-theory or epistemology---we have no trouble making the needed judgments.[467]

2.3.1 Intrapsychic conflict in relation to repression$_1$

It's one thing to *suspect* that something is the case; it's quite another to know it. The police *suspect* that Smith is the murderer. The evidence points to him (the bloody footprints leading to his front door step, the hair samples, the smudged fingerprints which are consistent with Smith's being the killer even though they don't prove it). But the police don't yet know that Smith is their man. For there are gaps in the evidence they have on him, as a result of which there are semi-reasonable ways of modeling the scenario that, if correct, would entail that Smith was not the murderer.

Bearing these facts in mind, suppose the following. Smith is your best friend. You owe him your life many times over. The police strongly suspect that he's the murderer. You're the detective who is assigned the case. Like your colleagues, you strongly suspect that he's guilty. But you aren't sure. And you don't want to be sure. So you choose not to exercise due diligence.

What we call "repression" involves failure to exercise due diligence. We suspect, but we don't know, which is how we want to it to keep it.

So the customary conception of repression is erroneous. The customary conception is given by assertion that ("CCR" is short for conventional conception of repression"):

> (CCR) the person in question has some propositional attitude *P* such that (a) he knowingly chooses to cease to be aware of it; i.e., he knowingly drives it into his unconscious; and such that (b) *P* is constantly struggling to break into consciousness, failing to do so only because the patient keeps it out.

Part (b) of CCR is indeed paradoxical. For, presumably, to keep *P* out of his consciousness, the patient must know what *P* is: otherwise he won't know what it is he's trying to keep out. But if he knows what it is he's trying to keep out, then he's conscious of it and it is *ipso facto* unrepressed.

But CCR doesn't correctly describe the nature of repression. It's one thing to *know* that your best friend is the Springfield Killer. It's another thing to suspect it but choose not to follow the relevant trail of evidence. Repression---or, rather, repression$_1$ (the subscript has again become relevant)---is a *failure* to do something; it is a failure to perform the operations by which one customarily generates one's beliefs; it isn't an operation performed on existing beliefs (or

propositional attitudes of any other kind). So repression$_1$ involves choosing *not* to learn something that, if learned and therefore known, one would wish that one had never learned, but only after it was too late.

While your parents are on vacation, let us suppose, you look through their bedroom and come across your mother's private journal. On the basis of its external appearance, along with the information that you've gleaned from a cursory glance at its interior, you know enough to know that the journal provides detailed information about your mother's sexual exploits over the last few years. You thus know that, *were you to read on*, you'd learn things about your mother that you know you don't want to know. So you choose not to read on. Your wish not to be traumatized overpowers your curiosity.

Repression$_1$ is to be understood along these lines. When we speak, inaccurately, of "repressed knowledge," we're referring (in at least some cases, viz., those that aren't cases of repression$_2$) to suspicions that we have chosen not to verify, since we suspect that, were we to do so, we'd be stuck with knowledge that we wish we didn't have.

Every case of repression$_1$ does, as CCR says, involve intrapsychic conflict. But, contrary to what CCR says, the conflict is not between unconscious knowledge of some truth *T* on the one hand, and conscious ignorance of *T* on the other. The conflict is between a wish to know the truth and a wish to *not* know the truth.

2.3.2 One reason why repression$_1$ often fails

One's agency only goes so far; there is much within one's mind that falls outside its scope. Among the various mental entities that lie beyond the reach of what it is for one to *choose* to occur are the aforementioned judgment-producing processes. So it is not within my power to completely suspend those processes.

And so far as it *is* within my power to do so, it is not in the direct, robust sense in which one can choose to move one's arm, but in the indirect and attenuated sense in which one can choose to become obese. One doesn't really choose to become obese; one chooses to do things that will redirect automatic processes in such a way that obesity is the inevitable result. Similarly, one can no more choose to stop making judgments than one choose to stop digesting food. One can choose courses of action that will have the effect of redirecting the automatically occurring stream of mental activity that yields judgments (e.g., one can choose activities that will flood that judgment-making faculty with information concerning the judgments that one wants to avoid making); and one can also choose to do things that will deplete that faculty of the energy it needs to function (e.g., one can do intense exercise for hours on end). One can also choose to paralyze or even permanently damage the physiological underpinnings of those judgments (e.g., one can get drunk, thereby paralyzing them, or one can get a lobotomy, thereby destroying them). But, unless one goes so far as to damage one's brain, that stream of activity will eventually start yielding the very judgments that you don't want it to.

2.3.3 The concept of anxiety in relation to repression$_1$

The anxiety that the subject feels when the forces of repression$_1$ break down isn't the result of some hitherto unconscious knowledge entering consciousness. It is the result one's own thought-processes, despite one's attempt to stymie them, following up on clues that it had good reason to believe would lead to a belief that one doesn't want to have.

"But if you already know that you don't want to have that belief, then you must already know what that belief is and must already have it." This is false. Suspicion isn't knowledge; and there's nothing paradoxical about not wanting to pursue suspicions that one suspects would, if pursued, lead to knowledge one would rather not have.

Also, one can know what *sort* of thing something is without knowing *which* thing it is. If I ask you to find a unique number *n* such that the square root of *n* is over 10 and such that two people whose middle names are "Robert" each own $n^{1/3}$ acres of land, and I prove to you, without identifying the number in question, that there is indeed exactly one such number, you know what *sort* of thing you're looking for without as of yet being able to identify any number that fits the description.

A story will make this clear. You are acquainted with Sally. You know that exactly one person has been stealing money from your wallet---but you don't know who it is. So, by virtue of knowing the description "person who has been stealing money from my wallet" to be uniquely satisfied, you have descriptive knowledge of Sally. But you don't know that Sally satisfies that description.

3.0 Repression$_2$

Whereas repression$_1$ involves *failing* to form a belief, repression$_2$ involves blinding oneself to an existing belief that one has already formed.

How is repression$_2$ possible? How does one repress a belief that has already been formed? And how does one *keep* such a belief repressed? To clarify the second question, how does one *keep* a belief that one doesn't want to enter consciousness from doing so? For one must be conscious of the enemy hammering at the walls of one's fortress if one is to take measures to keep him out. But, in this context, one cannot be conscious of the enemy, since what one is conscious of *ipso facto* isn't repressed.

3.1 A necessary condition for repression$_2$

For a propositional attitude to be a candidate for repression$_2$, it must be a source of *conflict* within one's own psyche. It isn't enough that it be a source of *pain.* One often fails to repress$_2$ one's most painful memories. (Holocaust survivors remember what they've been through.) People typically choose to *sup*press painful memories. In other words, they choose not to think about them. But suppression isn't repression$_2$.

One represses$_2$ knowledge that one has about one's husband, because one's life is so structured that it would be severely disrupted were the truths thereby known to be admitted. One likes many aspects of one's life; it may well be that, despite the ugly truths in question about one's husband, one believes (perhaps rightly) that being married to him really is better than not being married to him. But, since one couldn't stomach marriage to him if one admitted to oneself that he was a child-molesting ghoul, one represses$_2$ that knowledge.

And one may be right to. Repression$_2$ isn't always a bad thing. In many contexts, it's the optimal thing. It may be that Sally's quality of life would plummet if she left her child-molesting husband; it may be that, were she to leave him, her children would be in danger. Since she couldn't be with him if she consciously knew that he was child-molester, she's well advised to repress$_2$ that knowledge, at least for the time being.

A person who never repressed$_2$ anything would quickly go mad. Repressions$_2$ are necessary. So how is repression$_2$ bad? First of all, repressions$_2$ outlive their usefulness. The circumstances that prompts one to repress$_2$ a given belief may well be replaced by circumstances in which one's consciously having that belief is vital to one's wellbeing. (In this context, take "belief" to stand for any propositional attitude.) But, repression$_2$ being what it is, one will have no idea what it is that one is supposed to be aware of. In fact, one is unlikely to know that one has represssed$_2$ *anything* important to one's welfare.

Also, repression$_2$ isn't a one-off affair; it is an on-going process. Whenever one comes across information that confirms the repressed$_2$ belief, one must either repress$_2$ one's knowledge of that information or one must convince oneself of the cogency of some spurious interpretation of that information. Since

these cover-ups must themselves be covered up, more knowledge ends up being repressed$_2$ and more spurious rationalizations are accepted. Each instance of repression$_2$ creates a hole in one's body of knowledge that one must explain away. Since these explanations cannot be genuine, and since each one incentivizes the creation of more bogus, anomaly-creating explanations, the need to plug up these holes disposes one to devalue truth and to hyper-value the methods of reasoning whose legitimacy these faux-explanations falsely assume. The result is a loss of both ratiocinative and ethical integrity.

So, while repressions$_2$ are not necessarily bad things, sometimes being indispensable to one's well-being, they tend to stick around long after they've served any legitimate purpose; and during that time, they tend to spread and, as a result, to eat away at one's psychological architecture.

3.2 A repression$_2$-conducive factor

One can know something without that thing's being *obvious* to one. I know that there is a war in Iraq. But it's not *obvious* to me the way that it would be if I were there witnessing it. It's harder to repress$_2$ knowledge of the obvious than of the non-obvious. If your knowledge that your husband is a serial killer is based on inferences, even air-tight ones, it will be easier to repress$_2$ it than it would be if it were based on personal observations of him killing people.

That said, people *can* repress$_2$ knowledge of truths that are as obvious as truths can be. But one cannot repress$_2$ such knowledge without, in effect, telling oneself that one isn't capable of knowing the truth. Since the purpose of repression$_2$ is the elimination of conflict within oneself, and since a complete loss of faith in one's ratiocinative ability usually, but not always, creates more such conflict than it eliminates, repression$_2$ tends to target knowledge of what is not *totally* obvious. In most cases of repression$_2$, there must be what lawyers refer to as "plausible deniability."

3.2.1 A related point

The things that one represses$_2$ tend to be things that, although one knows them to be true, involve one's making judgment-calls. Until your husband's psychopathy assumes some particularly brazen form, any knowledge that you have of it involves your putting together various bits of data (e.g., various odd comments he's made, various murders that you hear about on the news, his unexplained absences during the times that when the murders occurred, his fitting the descriptions given to the police of the murderer). But, having as it does a basis in some many pieces of data, your subsequent knowledge that your husband is the murderer is shaky and is thus vulnerable to repression$_2$, in a way that strictly sense-based knowledge is not.

"But if one's 'knowledge' of some truth is shaky," it will be said, "then one isn't convinced of that truth. And if one isn't convinced of it, one doesn't have knowledge of it. One only has suspicions."

There are different reasons why one may lack confidence in one of one's beliefs. It may be that one has not, as of yet, encountered adequate evidence for that belief. Or it may be that, even though one has encountered adequate evidence and even though, consequently, that belief *would* ordinarily be firm, one wants that belief to be false. If a belief concerns something personal, one's standard of proof tends to be ridiculously high. What it would take to convince you that somebody you're not married to is cheating on his wife is miniscule to what it would take to convince you that your husband is cheating on you.

To take a different example: the amount of evidence that favors the theory of evolution is simply enormous. But because that theory says, in effect, that we're orangutans whose lives have no meaning, we won't judge it worthy of acceptance until it's met absurdly high standards of evidence. I have no doubt that people have *always* known that we evolved from other life forms. Given the anatomical, behavioral, and psychological similarities between us, on the one hand, and members of other mammalian species, on the other, it's inconceivable that we aren't related to them. Denying that orangutans are our biological kin is like denying that the one person in the room with grape jelly and crumbs around his mouth is the one who ate your jelly-doughnut.

Of course, religious people say that God created us. But, if so, why did he choose to make us look so much like apes, and why did he choose to stock our minds with ignominious urges that constantly show us that those external

similarities are but representations of internal ones?

This brings us to another illustration of the principle in question: it’s often said that Freud’s theories aren’t warranted by the empirical data. If you actually read what he says and then, bearing what you read in mind, study a single person’s behavior for a period of around 10 minutes, it becomes clear that Freud’s theories are theoretically extraordinarily conservative; it also becomes clear that he was simply making explicit the principles that we ordinarily use when making intuitive judgments about other people. People don’t want to be told that they lie to themselves. People don’t want to part with rationalizations; they don’t want to face ugly realities. They’d rather be a king in their dreams than a janitor in reality. So it’s not hard to see why it is often dismissed, despite its staggering explanatory power and despite the extraordinarily high degree to which his claims accord with the principles that underlie our judgments. We'll state Freud's principle claims right now.

3.2.2 Freud's Principle Claims

Here is a one-paragraph outline of Freud's primary contentions:

> Nobody is perfectly transparent to himself. Nobody is conscious of everything that goes on in his mind. There are different reasons why one may not be conscious of given constituent of one's own mind. In some cases, one can readily become conscious of a constituent of one's mind of which one was previously unconscious. For example, let's suppose that, since the time you woke up this morning, you haven't been conscious of your belief that 2+2=4. You just *became* conscious of it. Given how effortless this transition was, it didn't involve your resisting any forces within yourself. It was unconscious because, until just a moment ago, you lacked a positive incentive to pay attention to it----*not* because you had a positive incentive to turn a blind eye to it.
>
> But in other cases, a given person may be unaware of a given constituent of his mind for the reason that he doesn't *want* to be aware of it. He finds it so distressing that he chooses to turn a blind eye to it---to *repress* it, in other words. In repressing that entity, the subject didn't destroy it; he merely blinded himself to it. The subject wouldn't have found that entity important enough to repress if he hadn't known it to have a profound bearing on how he should live and, therefore, on what he should aspire to do. Consequently, the desires to which it gives rise are necessarily especially intense ones, and the subject will experience intolerable frustration until they are discharged.
>
> Such a desire cannot be discharged unless it employs a subterfuge of some kind, given that it cannot possibly be the subject's conscious intention to gratify it. Such subterfuges are always of one of two types.
>
> Type #1: The desire in question can disguise itself as a desire that the subject consciously *does* wish to act on. An example: the subject detests his ailing mother. But he refuses

> to believe that he has anything but love for her. His detestation for her presents itself to his consciousness as a fanatical devotion to her being safe from both microbial and human predators. So he bathes her in scalding hot water and refuses to let her leave the house. In this way, his repressed hatred of her is discharged in the form of behavior that the subject believes to be motivated entirely by unalloyed concern for her well-being.
>
> Type #2: It can be expressed as a symptom. An example: Having repressed her complete detestation of her husband, a young woman finds herself gripped by a compulsive desire to do housework whenever her betrothed wishes to engage her in acts of physical intimacy.

So far as the opposition to these contentions has been given an argumentative foundation, the arguments in question are pitifully weak and involve regressions to extremely crude and, in some cases, demonstrably false positions. “We’re computers” (Fodor). “There is no consciousness” (Dennett). “There are no minds” (Quine). “The mind is completely transparent to itself” (Searle). Nobody believes or doubts or hopes for anything; terms referring to propositional attitudes (e.g. "hope," "believe," "intend") belong to ‘folk-psychology’, which is no more accurate than folk-physics” (Paul and Patricia Churchland). “We’re better off trying to understand Baryshnikov's skill as a dancer in terms of fluid levels in his ears than his ability to upload, process, and distribute information in ways that ensure that his movements are appropriately coordinated.” (Searle).

These willful denials of the obvious are often embedded in larger philosophical projects that are nothing short of risible, e.g. the Dretske-Fodor view that, details aside, there is no difference between my concept of a given a person and a photograph of a given person.

And yet all of these flights into inanity are touted as “scientific” and “rigorous.” Why? “Because”, we are told, “in denying the existence of the mental, we’re simply completing the process of ceasing to understand nature in animistic terms and we’re thereby completing the victory of science over non-science.”

But this argument is no good, and it's clear why. While it's certainly a good idea not to impute minds to stones, it's obviously not such a good idea *not* to impute them to people. So while it’s a good idea not to understand the world

projectively, we must remember that the projector itself cannot be a projection.

3.3 Intrapsychic conflict in relation to repression$_2$: some preliminaries (continued)

A story, based on the remarks made in Sections 3.0-3.2.1., will help us clarify and develop the points we've made about self-deception and repression$_2$. You discover that your husband's a psychopath. You don't see him killing anyone. But within the limits set by that fact, you are as sure of it as a rational person could be. Since this knowledge (or near-knowledge) cannot be reconciled with your life, at least not in its current form, and thus creates intolerable conflict within your psyche, you choose to repress it.

Here is what repressing$_2$ this knowledge involves. First of all, when taken with other propositions whose truth we may reasonably assume you to accept, the proposition:

> (1) My husband is a psychopathic ghoul all of whose protestations of love were fraudulent, as was everything else about him; which, given what it is to be a psychopath along with the well-known (to people interested in the subject) symptoms of psychopathy (glibness, specific lack of insight, etc.)

entails the much more generic proposition:

> (2) There is some property *P* such that my husband has *P* and such that, by virtue of having *P*, my husband is a worse person than he would otherwise be; and a person's having *P* involves his doing the sorts of things that my husband does that trouble me (e.g., it involves his smoothly lying, his vacillating between inexplicable rage and glib charm, the contrast between how he acts and who he claims to be, his inability to authentic his claim that he was an admiral in the Navy).

By virtue of knowing (1), therefore, you also know (2). So, during the fleeting moment when you consciously know (1), you *also* know (2). Let's henceforth refer to that fleeting moment as "FM."

(2), unlike (1), is something that you can keep *in* consciousness without having to overhaul your entire life. For (2), unlike (1), is sufficiently vague that your knowledge of it won't make you want to retch at the mere sight of your husband. During FM, you realize that you wish you hadn't discovered (1). But, as you know, if you *simply* decide not to think about (1)---if you suppress, but do not repress, that knowledge---you won't be able to live with yourself. You know that you must bury your knowledge of (1). But you also know that you'll be confronted with the same facts---the strange behaviors, the news reports, etc.---that led you to know (1) in the first place; and you know that, unless you make some drastic change, those same facts will lead you right back to the same ugly place.

So you make a decision. You know (2). And you know that you know it. With this knowledge in mind, you resolve that, henceforth, when confronted with the sort of datum that that led you to believe (1), you'll model it in a way that is consistent with (2) but that doesn't coincide with (1) or with anything else that would be equally disruptive to your current situation.

Before we can move forward, we must recall some facts that we've already discussed about propositional attitudes, and we must point out a couple of new ones. First of all, propositional attitudes are not, as we've made clear, *constituents* of consciousness. For a propositional attitude to be "conscious" is either for one to be conscious *of* it or for it to be such that one can readily become conscious of it. Your knowledge that 7 + 3 = 10 is conscious in the sense that you can readily become conscious of it. You weren't conscious of it a few seconds ago, however. But, even then, it was "conscious" because it was readily made into an *object* of consciousness. (Freud (1915, 1921) describes such propositional attitudes as "pre-conscious.") So for your knowledge of (1) to be "conscious" is for it to be a ready *object* of consciousness.

3.4 Intrapsychic conflict in relation to repression$_2$: some final preliminaries

There are some other relevant facts. Beliefs are inert except to the extent that one draws consequences from them and, more generally, that they are allowed to beget or otherwise engage other beliefs. The making of inferences at least sometimes involves a deliberate, conscious effort on one's part. But some inferences are made involuntarily. Insights often just *occur* to one; but those insights are based on rational consideration of the relevant evidence. But, even though (so I believe) many an inference is made involuntarily, one has at least some control over what sorts of inferences one's mind makes. So, given some belief *B* that one has, one can so structure one's life that it is very unlikely that one will (consciously) make inferences from *B* that one would otherwise make. [468] Thus, one can so structure one's life that a belief, though still conscious, is (at the level of consciousness[469]) isolated---prevented from engaging the bodies of belief that it would otherwise engage and prevented from begetting the beliefs that it would otherwise beget. Such a belief is conscious in some formal sense; but, at the level of consciousness, it has become inert, like one's knowledge of some random, uninteresting fact.

By de-operationalizing one's acceptance of a given fact, one can cease to have anything more than an abstract, formal understanding of what that fact amounts to. I know that there are people who are being tortured. But, because I have not witnessed a torture session, I don't know what the proposition *there are people who are being tortured* amounts to as well as I would if *ceteris paribus* I *had* witnessed a torture session. So even though I am in some technical sense no less aware of that truth than somebody who has first-hand knowledge of what torture-sessions involve, my awareness of that fact doesn't mean very much to me, whereas his awareness of it means a great deal to him.

There is one last preliminary fact. Your knowledge of (1) is not derived from some one sense-perception; (1) isn't the sort of truth that can be known in such a direct way. It embodies a case of ***inference to the best explanation***. (See Chapters 10 and 11 for a definition of this term and an analysis of concept expressed by it.) That doesn't mean that it isn't knowledge. (Unless knowledge can be acquired through inferences to the best explanation, then what we actually know is only an infinitesimally small part of what we think we know.) But it does mean that it embodies an attempt on your part (a successful one, in fact) to model various pieces of information.

There are different ways to model that body of data. (This is not a peculiarity of the data involved in this particular case. Given a body of data, it is more likely than not than there is more than one reasonable, or at least semi-reasonable, way to model it.) [470] In this case, so we will suppose, (1) is the best such model. But it isn't, so we will also suppose, the only one that isn't totally implausible. The following proposition constitutes an alternative way of modeling that data:

(3) Given that mental activity is governed by brain-activity, it follows that there are (possible, if not actual) physiological conditions whose symptoms include behavior similar to my husbands, accompanied by bouts of forgetfulness. (Not being a physician, I don't know the specifics.) And his having such a condition would explain how it is that, without being intentionally dishonest, he often says things that are false, and otherwise behaves in substandard ways.

(3) is not a spectacularly good model of the data. But it's far from being *obviously* false. And this, for reasons that we'll now discuss, is enough to repress your knowledge of (1).

3.5 Intrapsychic conflict in relation to repression$_2$: a conjecture as to what it involves

In light of these points, I put forth the following conjecture. During *FM* (the fleeting moment when your knowledge of (1) was conscious), you resolve that, on the occasions that you become conscious (not *pre*-conscious, but *actively* conscious; i.e., conscious in the sense in which you are now conscious of reading these very words) of your acceptance of (1), you will take the following steps. First, you'll isolate that belief. You will, to the best of your ability, refrain from drawing inferences from it (or from conjunctive propositions of which it is one of the conjuncts). You won't voluntarily make any inferences from it; and if you find yourself having made one *in*voluntarily, you'll do your best to refrain from pursuing it. You'll encapsulate this knowledge.

This will neutralize your belief in (1). First, for reasons given a moment ago, it will blind you to corollaries of (1) of which you'd otherwise be aware, and (1) will start to seem relatively innocuous as a result. A story will help us develop this point.

You discover that your husband has done *X*, *Y*, and *Z*. (Some jewelry is missing from your drawer. Your husband is wearing new clothes, even though he has no money. And so forth.) Since you've decided not to draw inferences from (1), you'll be more likely than you would otherwise be to take these on a case-by-case basis. Inevitably, you'll sometimes connect these curious phenomena to your knowledge of (1). But, just as inevitably, bits of knowledge that you have that would otherwise congregate will remain isolated from one another. As a result, your appreciation of just what it is that (1) entails will remain superficial; you'll be slow to recognize its consequences for your own welfare, and to recognize the extent to which your husband's condition is impoverishing your life.

Having decided to isolate (1) you'll resolve that, whenever you encounter data that clearly confirms (1), you'll try to look at it from the viewpoint of somebody who believes that (3) is correct. Of course, since you are not such a person, this will, at first, be a kind of intellectual exercise for you: you'll still know that (1), not (3), is the right way to model the data. But by engaging in this exercise, you are ensuring that energy and thought that would otherwise deepen your understanding of the consequences of (1) don't have that effect at all: Those energies are now on a fool's errand. Also, this way of responding to such data will soon become automatic, for reasons similar to those for which, if an

actor rehearses his lines often enough, it becomes automatic, even though it wasn't before, for him to say *Y* if he just said *X*.

This response-mechanism will further isolate your knowledge of (1). That knowledge will, at least at the level of consciousness, become more and more inconsequential. One cannot in any direct way choose what to believe. (One can choose what to believe only in the sense in which one can choose to gain weight. One can choose to do things that one knows will add mass to one's body; one cannot choose to gain weight in the way in which one can choose to move one's arm.) But, within limits, one *can* choose which beliefs to make operational and which ones to make non-operational; one *can* choose to do things that will make a given belief impotent and irrelevant. I'm an actor; I'm mediocre at it. I know that fact. I cannot directly choose to cease to have that belief. But I *can* choose to make it irrelevant; for I can choose an occupation in which my lack of acting ability is irrelevant. Similarly, you can, though to a lesser extent, choose to adopt certain practices that make your knowledge of (1) irrelevant or, in any case, make it less relevant than it would otherwise be.

In this way, your acceptance of (1) becomes, at the conscious level, a mere shadow of what it would otherwise have been and of what it once was. It becomes like an old officer who, having become senile, is no longer allowed to command troops but who, because the higher-ups feel sorry for him, is allowed to keep his title.

3.6 Why repressions are never complete successes

Most data-synthesis occurs below the threshold of consciousness. When you walk into a room, you automatically upload a mass of information and you automatically make many inferences from that data. The uploading and synthesizing of information doesn't need your conscious cooperation. If it did, you'd know very little indeed. (In fact, it doesn't involve *your* cooperation at all, since most of it happens sub-personally. As we've seen, minds are more basic than selves. Your self is an agency that regulates existing mental activity, most, if not all, of which consists in the uploading and synthesizing of information. So, speaking both logically and chronologically, the *sub*-personal gathering and processing of information is more fundamental than the personal gathering and uploading and processing of information.)

It follows, and is clear on independent grounds, that your decisions as to what goes on in your own mind are of limited scope. Those decisions affect how well the conscious (and pre-conscious) strata of the mind communicate with other strata. They don't affect in any direct way affect what goes on in other, deeper strata. So, even if you've sincerely resolved to make no effort to interpret aberrant behaviors on your husband's part, that resolution fails to shut off a great deal of the ratiocinative activity that such a discovery automatically precipitates.

Armed with this information, let us return to our story. You have turned your knowledge of (1) into a dead husk, at least as far as its engagements with other denizens of consciousness are concerned. You have deliberately structured your life in a way that drains your knowledge of (1) of much of its importance. So far as it was in your power to do so, you have chosen to refrain from asking questions that you'd otherwise ask and, therefore, from making inferences that you'd otherwise make. You have thus so eviscerated your knowledge of (1), *qua* denizen of consciousness, that it might as well not even be conscious. You have thus *repressed it.*

But we have a problem. As we made clear in the second to last paragraph, you weren't able to shut off all of the processes that link that knowledge with other beliefs of yours. You could redirect some of those processes and contaminate others; and maybe you could altogether shut off some of them. But you failed to shut off, or even adulterate, much of that activity. And you didn't really cease to accept (1); rather, you ceased to treat that belief's manifestations in consciousness in the way in which you treat those of other beliefs. But, just as you cannot typically uninstall a computer program by dragging the desktop icon

for it into the desktop garbage can, you cannot "uninstall" a belief by tinkering with its representations in consciousness. Beliefs are enduring structures. Your belief that 1 + 1 = 2 has been around a long time; it does a lot of work for you even when you're not consciously deploying it. If you really stopped knowing that 1 + 1 = 2, your way of dealing with the world would suddenly change in a radical way; and the changes would pervade practically every field of endeavor, not just those in which you are required to consciously think that 1 + 1 = 2. That belief isn't like a mood or any of the other will-o-the-wisps that populate consciousness. And, like any other enduring structure that is constitutive of one's metabolism, one has, at most, partial control over its fortunes; and, so far as one does have such control, it tends to be of the indirect kind previously discussed. Given any other propositional attitude, the same thing *mutatis mutandis* is true of it.

Thus, the conscious representative of your belief in (1) is an innocuous shell. It has been *repressed.* But it still lives on, and it "presses for discharge" to use Freud's term. Since, however, you no longer acknowledge its existence, it cannot be discharged through intentional behavior on your part; it must, without your knowing it, insinuate itself into your intentions (e.g., you displace your repressed rage at your father by being mean to your students) or it must be expressed as a symptom (e.g., when you come into contact with situations that remind of your rage towards him, you become crippled by anxiety).

Whatever it is that you've repressed, in order to *keep* it repressed, you must be able to identify it. But if you can identify it, then you're conscious of it, and it's not repressed. So how can it be repressed? We'll now answer that question.

3.7 Intrapsychic conflict in relation to repression$_2$: conclusion

I submit that, if completed in the following way, our story describes one way in which repression$_2$ can occur.

During *FM*, you learned that (2) is the case, and you consciously choose to retain that knowledge. This enables you to have your cake and eat it too. (1) and (2) hold in virtue of the same, ugly fact. But (1) is much more specific than (2), and an acceptance of (1) is commensurately less likely than an acceptance of (2) to upset an existing psychological integration. At the same time, by virtue of knowing (2), you enable yourself to continue to be aware of an ugly an undeniable fact, and you've thus spared yourself a constant, difficult struggle to not see the obvious. (2) is enough like (1) that, by consciously accepting (2), one's conscious disposition isn't completely antagonistic to the still rational side of you that urges you to accept (1). So, by accepting (2), you're not locked into a draining struggle with the forces of rationality and truthfulness within oneself. At the same time, by accepting (2), you are neutralizing the message of which those forces are trying to apprise you, and you are thus spared some painful realizations as well as the labor-intensive changes to oneself that they demand.

3.8 Summing up

We can't choose to destroy the beliefs we don't want to have. But we can partly de-operationalize them. We can choose not to make any conscious effort to delineate their consequences. We can choose to immerse ourselves in activities or lifestyle that don't permit us to pay attention to them. And, though belief is no more voluntary than disbelief, we can also choose to look for reasons to have beliefs that are opposed to those that we wish we didn't have. We can look for such reasons without having much faith that we'll ever find them, just as a lawyer can build a case for a client he knows to be guilty. And the mere fact that we're looking for them means that the belief we wish to destroy no longer fits smoothly into our psychological architecture, since it no longer has the support of the psychological structures implicated in the search. So even though we can't directly choose our beliefs, we can choose to do things that make it harder to pay attention to a given belief than to withdraw our attention from it. The result may be that it becomes *so* hard to attend to that belief that it becomes one of which we becomes altogether unconscious.

3.9 Rationalizing rationalization

People sometimes repress their knowledge of ugly truths about themselves, and they must rationalize in order to prevent the knowledge they've repressed from re-entering consciousness. Unfortunately, the science of psychology has itself become little more than a means by which people rationalize ludicrous attempts to repress their knowledge of these obvious facts.

Many claim to believe that there is no unconscious mental activity and *a fortiori* that self-deception, repression, etc. are all artifacts of dated psychoanalytic theories; many claim that the concepts of repression, etc. are internal to a discipline (psychoanalysis) that have been superseded by 'rigorous' studies of the physiological underpinnings of thought. The thoroughly hypocritical and self-deceptive character of such pronouncements is evident when one looks at the judgments and conduct of these naysayers. Not a single one of those judgments concern the physiological underpinnings of a person's psychology; they concern that person's psychology. And, though this could technically be explained in terms of the fact that people can't routinely subject others to PET-scans, those PET-scans would be of no psychological interest---to anyone---unless it were shown that the findings of those PET-scans correlated with independently known facts about psychology. The giant discoloration on Smith's PET-scan would be of no interest to psychology were it not known *independently* of the results of that PET-scan that Smith's mentation was aberrant---that Smith thought that he was E.T., or some such, and was therefore mentally ill. Were it discovered that mental activity were mediated by the gall-bladder, the results of PET-scans would immediately cease to be of any psychological interest; for they would no longer be correlated with independently known, and independently explained, psychological data. Which shows that brain-science, to the extent that it's of any relevance to psychology, is parasitic on psychological knowledge.

Most of those guilty of this hypocrisy provide no arguments for their view, other than empty, boiler-plate repetitions of the slogan that "science" studies brains and molecules, not intangibles such as emotions and thoughts.

As for the idea that emotions, thoughts, etc. are too intangible to be worthy of scientific study---that position is decidedly implausible. Whatever we know about the external world—and, therefore, about brains, etc.---we know it through perceptions and thoughts and thus through such alleged intangibles. So while brains and neurons may be ontologically more basic than thoughts,

thoughts and perceptions are epistemologically more basic than neurons and brains. Indeed, the term “brain” is, believe it or not, a theoretical term; in characterizing something as a “brain," you are ascribing to it various functions that, without engaging in a great deal of theory-soaked extrapolating, you cannot know it to discharge. “Brain” is a theoretical term that is deeply entrenched in a well-confirmed body of theory. (Hence the *illusion* that it is a non-theoretical term.) So, speaking epistemologically, these psychological “intangibles” are much more tangible, and their existence infinitely less dubitable, than “tangible” things such as brains.

But even though many rail against the inevitable encroachment on their lives of psychodynamic truths that they’d very much like to stave off—even though contemporary professional psychologists, having invested their lives in the bankrupt and tautologously false idea that psychology ought to be studied non-psychologically, are especially shrill in their antipathy towards any attitude towards the psychodynamic approach that isn’t one of unalloyed, unquestioning antipathy---very few people have tried to produce *bona fide*reasons to reject the tenets of such an approach.

I’ve heard more than one professor of psychology explicitly state that the study of emotions must be replaced by the study of neurons. Why? “Because,” they say, “the study of emotions is itself an emotional enterprise.” This is like saying that thoughts about plants are themselves plants. That anyone of even minimal intelligence should commit such a crass and grievous ratiocinative gaffe is incomprehensible. But it becomes comprehensible if we adopt the hypothesis that, within the mind of a person who advocates this view, there is something with which such discussions resonate uncomfortably and to which that person wants to continue to remain blind. Thus, the denial of emotions is itself a thoroughly emotional position, and the denial of repression is itself an instance of repression.

Those who deny the existence of emotions and repression say that, in so doing, they’re being “scientific.” But it’s hard to see how such dishonesty and obtuseness could deserve to be so characterized. People don’t see themselves for what they are. Is it scientific to deny this? In some cases, lack of self-knowledge has the same causes as other forms of ignorance: lack of ratiocinative ability, lack of effort, lack of raw data. But given the empirical fact that hyper-intelligent people are quite as prone to not seeing themselves for who they are as their less gifted brethren, it is *de rigueur,* from the viewpoint of scientific methodology, to suppose another factor to be operative. And it’s pretty clear

what this other fact is: People don’t *want* to see themselves for who they are; so they don’t. Given the obvious and, outside of the discipline of psychology, universally acknowledged fact that people deceive themselves, there’s no coherent way to deny that repression is not only possible, but actual, and not only actual, but integral to the human condition.

3.4.4 Sartre on repression and the unconscious

Jean-Paul Sartre held there is no such thing as unconscious mental activity and *a fortiori* no such thing as repression. His position was that the mind is completely and utterly transparent to itself, it being in the nature of mentationthat a given mind should have immediate and complete access to all of its own contents. But unlike the previously mentioned authors, Sartre argued for his position. And his argument is initially persuasive:

> (SA)[471] One must know, said Sartre, what it is that one is trying to keep out of consciousness *if* one is to succeed in doing so. But so far as one has that knowledge, he continued, one is conscious of the very thing that one is trying to keep out of consciousness; and so far as one is conscious of thing, it isn't repressed.

Given this argument, there are two positions one might take:

> (1) Sartre is right and his reasoning is cogent.
>
> (2) Sartre is wrong. So his argument must be fallacious and we must pinpoint the fallacy in it.

In my book *Conceptual Atomism and the Computational Theory of Mind*, I try to identify an error in Sartre's reasoning. In this chapter, I have developed what I said in that book. But I must make it clear that, although I have no doubt that people do deceive themselves, repress conflicting engendering beliefs and sentiments, etc., I am not convinced that:

> (i) My attempts to show the concepts of self-deception (etc.) to be coherent were entirely successful.

And supposing for argument's sake that I have shown those concepts to be coherent, I am not convinced that:

> (ii) The specific forms I have shown them to be *capable* of

taking are the forms that they *in fact* take.

I've identified *one* possible model (or, rather, two such models: one for repressions$_1$ and one for repression$_2$). But I don't know that I've identified models that correctly identify the forms that those phenomena *do* take. Still, I am convinced that there are elements of truth in those models. And I think it is a philosophical task of the utmost urgency to make it crystal clear what is wrong with Sartre's powerful argument and also to make it crystal clear what exactly self-deception and repression involve. Unfortunately, philosophers have neglected to do this; which, so far as it's had any effect on psychology, has encouraged it to go down the blind allies of scientism, methodological incoherence, and (cf. the view now widely endorsed by theoretical psychologists) the obviously false views that constitute contemporary theoretical psychology.

Chapter 27 Psychopathy, Sociopathy, and Non-Psychopathic, Non-Sociopathic Criminality

1.0 some preliminary points concerning the concept of mind

A mind is a collection of structures such that

> (i) those structures create and direct mental activity;

and such that

> (ii) if X_1 and X_n are any two of those structures, one of two conditions is satisfied: (i) at least one of them can directly transmit the information stored in it to the other faculty; (ii) there is a set of faculties $X_1, X_2 \ldots X_n$ such that, for each i ($1 \leq i \leq n$), X_i, can directly transmit the information it bears to X_{i+1}.

(i) is self-explanatory; but (ii) requires clarification. In order for you to acquire information as to what I believe (or hope or fear, etc.), one of three conditions must be met:

> (a) I must engage in overt bodily behaviors (e.g., laughing, throwing furniture around) that you can interpret as indicative of my having a certain belief (or whatnot).
>
> (b) I must engage in behaviors that that result in ink-deposits, bursts of noise, or otherwise publicly knowable entities, that you can interpret as being thus indicative.
>
> (c) I must engage behaviors of the sort described in (a) and (b) that, although you do not personally witness them, initiate a testimonial-sequence of which you are on the receiving end.

Thus, the exchange of information *between* minds involves non-mental intermediaries and, for that reason, is in some important sense "indirect."

But information-transmissions that occur *within* a mind is either direct or, if

it isn't direct, *consists* solely of direct information-transmissions (i.e. that consist of sequences of information-transmissions, each one of which is direct, even though, obviously, those direct transmissions jointly constitute an indirect transmission). I know that Jerry is a vicious and prolific murderer. All of a sudden, I see Jerry's grinning face pressed against my window. The result: I'm afraid. (I wouldn't have reacted with fear if my pre-existing belief had been to the effect that Jerry was a harmless prankster who, in his attempts to be funny, pressed his grinning face against people's windows.) Clearly, the information encoded in my knowledge about Jerry is "pooled" with the contents of my sense-perception, that being why I am struck by fear. That pooling occurs without anyone's engaging in any overt behaviors and, therefore, without anyone's employing any symbolic intermediaries. The one mental entity's contents *directly* merge (if that's the right word) with the others.

The fact that these two bodies of information can merge thus directly is not a *consequence* of their both belonging to the same mind: it is *constitutive* of their belonging to the same mind. Two mental entities whose contents are thus related *ipso facto* both "belong" to some one mind.

Of course, there are distinct mental entities that belong to some one mind that are not as directly related to each other as are the two just-mentioned mental entities. For example, none of the mental entities constitutive of my mind in 1987 is related in that way to any given one of my current mental contents. And there are mental entities in my mind right at this very moment that are not thus interrelated. But, given one of those 1987-mental entities, *how* are its contents pooled with those of a given one of my 2010 mental contents? Through overt behaviors on my (or somebody else's) part? No. Through symbolic intermediaries? No. The pooling is accomplished via a series of mental states $S_1 \ldots S_n$ such that S_1's contents directly merge with those of the aforementioned 1987 mental entity, such that S_n's contents directly merge with those of the aforementioned 2010 mental entity, and such that, for each i, the contents of S_i directly merge with those of S_{i+1}. The same thing *mutatis mutandis* is true of any two mental entities that *simultaneously* inhabit a given mind but that, perhaps because the one's position in that mind is remote from the other's, cannot pool their contents with the directness with which my knowledge of Jerry and my sense-perception of Jerry's pale, grinning face can pool *their* contents.

Of course, a given mental entity may fail to leave a trace of itself and may

thus fail to transmit its contents to other mental entities belonging to the same mind, and the contents of two mental entities that simultaneously belong to a given mind may fail to be pooled together. But *were such a pooling of contents to occur*, it would be through a series of mental states of the just-described kind; it would *not* be through non-symbolic intermediaries (e.g., overt behaviors).

It incoherent to hold that, underlying the various wasps that compose a swarm of wasps, there is some entity, this entity being the swarm *per se*, to which those various wasps *belong* but of which they are not constitutive. It is similarly incoherent to suppose that, underlying the contents of a given mind, there is some "self" that *has* those contents but is distinct from them. Were such a thing to exist, its presence wouldn't enable wrongly interrelated mental entities to belong to the same mind, and its absence wouldn't prevent appropriately interrelated ones from belonging to the same mind. Such a thing would do nothing and would therefore be nothing.

2.0 Desires vs. values

Bearing in mind the points just made, let *M* be some mind (notice that I did not use the word "self"), and let $S_1 \ldots S_n$ be the series of aggregates of mental entities constituting *M*.[472] In every respect but one, *M* is just like a human mind. Being housed in a suitably equipped body, *M* can see things, touch them, etc., and it can thus acquire information concerning the external world. Being rational, *M* has the ability to draw inferences from that information that enable it to gratify its desires. And, because the body housing it is in good working order, it can act on the intentions to which those inferences lead. But *M* values nothing. This is the one respect in which it differs from (most) human minds. *M* wants things, one of those things being to survive.

But *M*, we will suppose, values nothing. *M* sometimes *seems* to value things; for there are things that it wants that don't have any obvious survival-value and that don't relate in any obvious way to the gratification of its urges. But, in such cases, *M* believes the thing in question to conduce to the gratification of its desires; and so far as *M* appears to value it for its own sake, it is only because *M* isn't disclosing why it believes that thing to be thus conducive.

There is *a* sense in which *M* has a concept of self-improvement. If it knows that acquiring the ability to sing well will, for some reason or other, enable it to gratify some desire that it cannot otherwise gratify, *M* can resolve to learn to sing well. In general, if *M* knows that, in order to gratify some urge that it has, it has to change itself in such and such a way, *M* is able to resolve to make that change. But, beyond that, *M* cannot possibly have any concept of self-improvement. To improve oneself is to change oneself in such a way that one comes closer to satisfying some standard that one has that, for that reason, one judges to endow one with more value than one had previously. Given that, by hypothesis, *M* values nothing, it thus has no conception of self-improvement apart from the strictly prudential one just described.

It follows that all of *M*'s actions are really *re*actions. It has desires—it wants to eat, sleep, and so on. Changes in circumstances sometimes bring about changes in what *M* has to do to gratify these desires; and *M* is reasonably adept at making the needed changes. But, given that *M* values nothing, *M* cannot possibly aspire to make any changes to itself that aren't intended to gratify some base urge. So any given action on *M*'s part that isn't a direct attempt to gratify a

base urge is an indirect attempt to so. Each of *M*'s actions falls into one of two categories: either it's an attempt (possibly a successful one) to eat or sleep (or some such) or it's an attempt to do something that, in its judgment, it must to do in order to eat or sleep (or some such). *M*'s conduct is always reactive.

So far as *M* has a self, it does nothing other than mediate between, on the one hand, *M*'s urges and, on the other hand, *M*'s beliefs as to what, given the circumstances, must be done if those urges are to be gratified.

By virtue of having no values, *M* is necessarily incapable of understanding much that creatures with values, but that are otherwise comparable to *M*, are capable of understanding. *M* won't really have the concept of friendship. (*M* will have a concept of a strategic alliance, but not of friendship.) *M* won't understand why Smith chose a non-lucrative and difficult career path over a lucrative and easy one. *M* cannot possibly have any concept of intrinsic, as opposed to purely instrumental, value.

The essence of the human condition will elude *M*. He feels nothing that involves an ability to value something for its own sake. He can feel frustration (because his efforts to find food are unsuccessful); he can feel rage (towards those who stand between him and the successful execution of his machinations). But he doesn't love anybody; he doesn't even like anybody. And there is nothing towards which he aspires.

M has no concept of legitimacy. There are desirable outcomes and there are undesirable outcomes. But nothing is any more *right* than anything else.

2..1 Desires vs. values (continued)

Some other facts about *M* deserve mention. *M* values nothing and therefore doesn't value the truth, at least not for its own sake. He values the truth only to the extent that, in his judgment, knowledge of it will help him gratify his desires.

For purely strategic reasons, *M* may learn to 'hit the right notes'---to say the right things. In fact, since there isn't anything that he wants to do or, therefore, to say that he doesn't believe to be to his strategic advantage, he'll do his best to do and say the 'right' things; unlike actual people, whose deeds and utterances 'come from the heart' and, therefore, are often *not* strategic, *M*, though *in actuality* completely heartless, will strike many as having an exceptionally pure heart. *M* always says the *correct* thing, as opposed to the *human* thing. He says what he believes it to be to his practical advantage to say, given what he has learned about human nature from his examination of it.

It is from an outsider's perspective that *M* examines people. It is thus to be expected that *M*'s analysis of human nature will be hollow and off-the-mark. At the same time, *M*'s analysis will strike less acute observers as unusually cogent. And, ironically, this is precisely because *M* is so blind to how people really are. Being incapable of saying anything substantive about the human mind, *M* will be forced to produce an analysis that is *formally* good, a corollary being that his analysis will be free of the ragged edges and inconsistencies that vitiate analyses put forth by people who, unlike *M*, have access to far more psychological data than can possibly be embodied in a few well-manicured truisms.

Also, in attempting to describe the human mind, *M* will be his forced by his lack of psychological insight to give undue weight to what people say about themselves. What a person says about himself doesn't reflect who he is so much as it reflect who wants to be and who, depending how convinced he is by his own rationalizations, believes himself to be. Thus, *M*'s descriptions of people will be consonant with their self-serving rationalizations and congenial to their fragile emotional integrations. As a result, they will strike people as accurate and heartening. By contrast, somebody who is psychologically insightful will see people for who they are; and, depending on how much stock people put in their own narratives, his psychological analyses strike people as false or threatening and will be rejected either way.

2.2 Desires vs. values (continued)

A corollary of the fact that *M* doesn't value truth for its own sake is that he doesn't *negatively* value falsehood for its own sake. He has no inhibitions, other than purely prudential ones, to accepting anything, so long as he thinks that, by doing so, he stands to come out ahead.

Granting that most people are dishonest to some degree---a very high degree, in most cases---they value truth for its own sake. This doesn't necessarily mean that they'll tell the truth or that they'll (consciously) accept it when they see it. But it means that a high, non-prudential regard for truth is operative within them and, in *some* way or other, will affect what they say, do, and feel. If they lie, they'll feel guilty; and what they refuse to accept but believe deep down often comes out in their actions. (They refuse to believe that Smith is an evil man and have but kind words to say about him. But they just can't find it in themselves to go visit him in the hospital when he's sick, they forget to pick him up at the airport, etc.)

But it goes without saying that *much* of the time a person's high regard for truth is expressed in his telling the truth. I've heard many a police officer say that criminals *want* to tell you the truth. Freud said that what a person doesn't explicitly say 'oozes' (Freud's own term) out of him; he fidgets; he unwittingly spins what he believes to be purely fictitious yarns that, when the names are changed, are precisely what the individual in question is trying to keep a lid on. (I don't know whether Freud or the aforementioned police officers had psychopaths in mind when they made these statements. I suspect that they were thinking of non-psychopaths who were dissembling.)

The reason that truth-serum (sodium pentothal) works so well is that it lowers the inhibitions that one has to erect, and expend considerable energy maintaining, in order to prevent the truth from pouring out of one's mouth. People are fundamentally truthful---even those that lie 99% of the time. I personally believe that very little of what people say is completely sincere and that the pervasiveness of mendacity cannot be overstated. But that isn't because people *don't* value the truth for its own sake. It's because, in addition to valuing the truth, they value their own welfare; and their concern for their own welfare overpowers their wish to tell the truth. But, echoing what we just said, this doesn't mean that a non-psychopathic liar has successfully decommissioned his regard for the truth. His high valuation of it comes out in rearguard ways: he feels guilty; he volunteers at an orphanage (to alleviate his feelings of guilt); he

is an unremittingly faithful husband; and, most importantly, the (non-psychopathic) liar will, at least at some level, keep a mental record of the truth, this being, at least in this context, the most important difference between non-psychopaths, be they liars are or truth-tellers, and psychopaths.

3.0 These points in relation to psychopathy

M is what is known as a "psychopath." This is because, for *M*, what is true has no more (non-instrumental) value than what is untrue, and the essence of psychopathy lies is a complete failure to regard the truth as having any non-instrumental value. Whereas the non-psychopath keeps a mental log of the truth and, since he genuinely values the truth for its own sake, does so *even when he doesn't judge there to be any practical reason to know the truth*, *M* keeps a log of the truth *only* to the extent that, in his judgment, doing so is a practical necessity.

This fact about *M* leads to no end of trouble for him. It's very hard to know *which* truths will be of use and which one's won't. (In fact, given points made in the previous chapter concerning the limits to which, even in a deterministic universe, the future can be predicted, it is, technically speaking, *impossible* to be unerring in one's judgments as to which truths will have which consequences and, therefore, as to which truths are worth knowing and which ones are not.) So it's extremely easy to make false judgments as to which truths are, and which ones are not, worth logging.

By the same reasoning *mutatis mutandis*, it's hard (or, strictly speaking, impossible) to know exactly which falsehoods are worth regarding as truths---or, in any case, as operational truths---and which ones are not. Since *M*'s commitment to keeping the facts straight extends no further than his conviction that it's in his interest to keep the facts straight, and since it's hard, if not impossible, to know which facts are the ones that are relevant to one's designs or *a fortiori* exactly *how* they are thus relevant, *M* often has no mental record, or only a very distorted one, of the facts. These gaps *themselves* change what it is to *M*'s practical advantage to believe.

Since, by obvious extensions of the points just made, these gaps in *M*'s knowledge tend to spread, *M* is forced to have an extremely flexible relationship to the truth; *M*'s failure to see the truth as having any intrinsic (non-strategic) value has the consequence that the purely *instrumental* value that the truth has for *M* tends to shrink with the passage of time.

3.1 These points in relation to psychopathy (continued)

For these reasons, *M*'s "narrative"---what he tells both himself and others---isn't consistent. In fact, *M*'s narrative is as mercurial as the weather. *M* has no stable self-concept. *M* values nothing. Apart from his having certain base desires, there is nothing fixed about *M*'s personality. Consequently, *M*'s personality changes every time he judges there to be a reason to change the manner in which he attempts to gratify those desires. His personality thus changes every time he judges there to be a change in his circumstances that is relevant to the methods he must employ to gratify his desires.

This means that *M*'s personality is as fleeting as a summer breeze. And *this*, in its turn, means that *M* doesn't really have a personality. He doesn't really have a self. He has base desires; he has (mis)information as to how to gratify those desires; and he has intentions corresponding to that (mis)information. But that's all he has. He is whatever circumstances require him to be. So he is a Potemkin village---an empty suit, a mask covering, not a face, but a void. *M* is a mind, clearly, but *M* is not a self; for it no more makes sense to speak of a self that changes with every decision-relevant change in circumstances than it does to speak of a square circle. A self is, by definition, what remains, or tends to remain, constant throughout the changes in the circumstances in which one is operating.

In the last section of this book, I discuss some famous psychopaths.

3.2 Who does the psychopath fool?

Even though everyone lies, the non-psychopath *tends* to speak from the heart and thus to say what he believes, as opposed to what others want to hear. The psychopath has no tendency at all to say what he believes. For he has no heart and, therefore, no heart to speak from. Since, therefore, his only reasons for speaking are strategic, he says what people want to hear. Thus, even though he doesn't have a heart, the psychopath is more likely than others to strike people as having an exceptionally pure one .

But the psychopath doesn't fool everybody. He only fools only those who fall into one of two categories:

> (i) Those who simply lack the acumen to see through his chicanery

and

> (ii) Those who, because of special circumstances, have an emotional incentive *not* to see through it.

Among those falling into the second category are

> (a) Those who are themselves children or parents of psychopaths

and

> (b) Those who, having been brutalized in some way or other, possibly by some *other* psychopath, find much needed solace in his trite blandishments and thus have an emotional need to believe in the authenticity of their source.

Those not thus blinded will see hollow performances in the psychopath's noble deeds and empty clichés in his insights. In due course, those who see through the psychopath are likely to see that there isn't a person behind the mask, only a proto-person who believes in nothing, values nothing, and is

nothing, and who, so far as he isn't trying to convince himself and others otherwise, aspires only to gratify base animalistic urges and quasi-human derivatives thereof, such as a need (not a desire—a *need*) for adulation or wealth.

4.0 Sociopathy vs. psychopathy vs. non-sociopathic, non-psychopathic criminality

The terms “sociopath” and “psychopath” are frequently used. Unfortunately, most people, without knowing that they are doing so, tend to use each of those terms to refer to three very different classes of people:

(1) Those who don’t have consciences at all;

(2) Those who have consciences but whose consciences have been corrupted by rationalization;

and:

(3) Those who have consciences, and whose consciences have not been corrupted by rationalization (or, necessarily, by anything else) but who are evil.

The result is that these three conditions, each of which is both psychologically and ethically very different from each of the others, are lumped together.

People in Category 1 are *psychopaths.* Before we talk about what psychopathy is, we must talk about what it isn’t. Psychopathy must be distinguished from psychosis. Psychosis is the psychological disorganization that results when there is a lowering of the inhibitions that ordinarily limit the freedom with which different parts of the psyche exchange information. A possible consequence of psychosis, but not an inevitable one, is a ‘break from reality’: the psychotic ceases to be able to distinguish between reality and fantasy. But, it must be stressed, such breaks from reality (which may, or may not, include hallucinations), are a by-product of psychosis, not the essence of it. (Incidentally, the hallucinations in question are usually auditory, not visual.) Psychotics are as likely as non-psychotics to have consciences. So psychosis is not psychopathy. Having cleared this up, we will now clear up some widespread misconceptions about psychopathy.

When people of my nationality and approximate age hear the term “psychopath," they tend to think of Dr. Hannibal Lecter---the evil super-genius

in the move *Silence of the Lambs.* The character of Hannibal Lecter is fictional in every possible sense. There is no real person on whose character his is modeled; and, human psychology being what it is, there couldn't possibly be such a person. He is about as unrealistic a character as one could imagine. He bears no resemblance to psychopaths---or, really, to anyone else. There is nothing to be learned from the character of Hannibal Lecter about psychopathy, psychopathology, or (setting aside the fact he has certain extremely generic traits in common with actual human beings; e.g., he doesn't like being in a cage) human psychology generally.

A related point is that history's great evil-doers are seldom, if ever, psychopaths. Stalin, for example, was not a psychopath.[473] For reasons that we'll now discuss, psychopaths don't have the attention-span to do what Stalin did and they cannot experience hatred, or any other emotion, with the depth with which Stalin did.

One's conscience has many functions. In certain cases, under certain circumstances, *one* such function is to prevent one from acting in anti-social ways. But that isn't the only function of conscience or even the primary one. Non-psychopathic individuals hold themselves to standards. One's conscience is the part of one's mind that stores one's standards and that punishes one if one falls short of them. The standards in question may not be morally good ones. To make the same point, but in a way that doesn't prejudge any moral questions: those standards may be ones that, if complied with, hurt others more than they help them and (what is a related, but not a coincident, fact) tend to undermine the social institutions on which people's welfare depends.

But speaking from a psychological, as opposed to an ethical, viewpoint, there is all the difference in the world between, on the one hand, holding oneself to "bad" standards, whatever that means, and, on the other hand, holding oneself to *no* standards. Were it not for the standards to which we hold ourselves, we wouldn't *have* selves to begin with. Creatures that don't hold themselves to standards may have *minds*---but they don't selves.

To understand how selves can be different from minds, we must first state some basic truths relating to what minds are. The points about to be made (in Section 4.1) may initially seem purely digressive and irrelevant. But their relevance will soon become clear.

4.1 Why psychopaths are dysfunctional

M is a very pure example of psychopathy. *Actual* psychopaths tend to have *some* values. They like to tinker with cars, to build model airplanes, or whatnot. But the minds of actual psychopaths differ surprisingly little from *M*.

There is one respect in which actual psychopaths resemble *M* that deserves special mention: like *M*, they have an extremely flexible relationship to the truth. (Psychopaths are unusually likely to be able to pass lie-detector tests. This isn't because they are good liars. It's because----setting aside the rationalization-proof deliverances of their *current*, as opposed to their remembered, sense-perceptions---a psychopath's relationship to the truth is so elastic that there isn't anything that he believes. So the psychopath doesn't *have* to lie. For the psychopath, some beliefs are more useful than others; so, for him, there is *operational* truth and falsehood. But there is no such thing as truth and falsehood *simpliciter*. Thus, psychopaths don't have to lie. They can pass lie detectors because, to them, the distinctions between truth and falsehood, and therefore between honesty and fraud, are purely technical ones.

Since, consequently, a psychopath's grasp on the truth is so tenuous, he is very unlikely to succeed in life.[474] His fakery, braggadocio, and showmanship often make him appear in his youth and early adulthood to be success-bound; and sometimes he achieves success, at least in a narrow professional sense. But their careers tend to fall apart. The highly conditional nature of their commitment to the truth forces them to make decisions on the basis of an unduly constricted body of knowledge, the results being that they are likely to commit horrible gaffes that alienate them from institutions within which one has to operate to succeed.

Also, because their grasp of the truth is so feeble, they are prone to act impulsively and therefore to commit acts that, to any non-psychopath, would obviously have consequences for them of the most adverse kind. According to Hervey Cleckley, whose book *The Mask of Sanity* is undoubtedly the best one on the subject of psychopathy, they do *not* do this because of an unconscious sense of guilt. There is, Cleckley says, no independent evidence that any such force is operative in these people. My own experience strongly supports Cleckley's contention.

4.1.1 Why psychopaths are dysfunctional (continued)

A related point is that psychopaths are extraordinarily bad at learning from their mistakes. [475] They make the same mistakes over and over again. They always have a story. And those stories often seem cogent when taken in isolation from one another and when considered in abstraction of the psychopath's back story. But their stories---which, in my experience, they describe using words such as "cogent," "rigorous," and "scientific"---turn out to be models of flimsy reasoning from half-true premises; and they have nothing more than a veneer of cogency, which, on the rare occasions when you can pin them down, are easily penetrated.

Psychopaths thus bear no resemblance to Hannibal Lecter and other Hollywood villains who are often (mis)characterized as "psychopaths"; and they bear no resemblance to similarly mischaracterized, real life villains, such as Stalin and Mao.

Psychopaths are dysfunctional. They can pull off cons. But that is pretty much *all* they can do. They can put on a good show for a while. They are charming, at least at first; they are, or seem to be, witty; and they often strike one as exceedingly intelligent and learned.

But psychopaths are like a house of cards.[476] To borrow a simile used by Mary Aston in her novel *The Incredible Charlie Carewe*[477], which depicts a psychopath, psychopaths are to people what snake-skins are to snakes. For one thing, the proverbial charm of the psychopath's is doubly fraudulent. First of all, it hides their absolute indifference to others. Second, at the end of the day, they *aren't* charming. Or, rather, *they* aren't charming: it is the people from whom they plagiarize that are charming. More often than not, a psychopath's genuinely witty comments turn out to have been lifted. And, when a psychopath's attempts at charm don't involve plagiarism, it isn't witty at all, consisting as it does of stale clichés and curiously two-dimensional pseudo-apercus.

4.2 The psychopath's veneer of intelligence

These points *mutatis mutandis* hold of the psychopath's proverbial intelligence. Psychopaths are more likely than non-psychopaths to show-case what they know, and they're more likely than non-psychopaths to choose endeavors that will lead to results they can show-case. Given a particular field, the psychopath will learn just enough to convince non-experts that he's an expert in it. If he knows that a given person is an expert in a certain field, he'll avoid that topic when talking with him or he'll avoid that person altogether. What the psychopath lacks in actual insight, he makes up for in plagiarized insights. What he lacks in both actual and plagiarized insights, he makes up for in his delivery. Psychopaths are long on delivery and short on content. A lecture given by a psychopath is likely to strike those who attend it as profound and to strike those who read transcripts of it as curiously hollow.

I would like to quote from Robert Lindner's (1944) classic *Rebel Without a Cause* (no relation to the movie of the same name, which it predates):

> In the literature on [psychopathy] frequent mention is made of the "high" intelligence of psychopaths…However derived, these notions are misleading or at least insufficient to serve either as diagnostic aids or explanatory propositions….So far as intelligence *per se* is concerned…what is outstanding about psychopathy is not its arithmetic proportions but is its peculiar variety and design…The intelligence of the psychopath can be described only adjectivally and in terms of the [psychopath's] whole personality. Perfectly adapted to his needs in the same way as protective coloring is suited to the preservation of the life of an animal, all those psychological functions (thinking, understanding, imagining, remembering, etc.) which are held to be components of "intelligence" have in the psychopath superimposed on them an aura of shrewdness and secretive cunning, of calculating canniness. These elements not only serve psychopathic ends but effectively distort and divert all known measuring instruments, the rigid designs of which prevent their divinatory use with such individuals…A further striking feature of the intelligence of the psychopath, and one which appears only after long time acquaintance with such

individuals, is concerned with the amazing excess-cargo of uncoordinated and useless information they possess. Frequently one is misled by their typically encyclopaedic range into considering them persons of high intellect, even of culture. Penetration with time, however, discloses that, like the veneer of mahogany applied to inferior wood, this mass of "knowledge" is superficial and undigested; that it is free-floating, lacking the requisite elements of cohesiveness and relativity.[478]

4.3 The reverence that people have for psychopaths

The reverence that people have for these empty vessels has nothing to do with the all but non-existent content of what they're saying; it has instead to do with their charisma and braggadocio. They exude an unabashed self-confidence and authoritativeness. When discussing U.S. history, a non-psychopathic expert on U.S. history is likely to appear self-confident, because he really *is* self-confident. And when he's talking about non-standard analysis (supposing that he isn't also a mathematician), he's likely to lack self-confidence and, therefore, to *appear* to lack it.

A psychopath, on the other hand, *always* exudes self-confidence. He may be an expert on what he's talking about, or may know nothing about it at all. In each case, he will exude the same level of confidence. There will be no changes in mannerisms, no vocal modulations, no trace of anxiety that would indicate self-doubt or compensatory hubris. Regardless how confident he *ought* to be, he will *appear* utterly self-assured. He is likely to use histrionics to draw attention away from what he's actually saying. In general, psychopaths are confidence-men (con-men). They con by exuding confidence.

There are two doctrines to which all but one or two of the many psychopathic professors I've known were drawn: Marxism and Cultural Relativism. What these doctrines have in common is that they deny there is such a thing as truth. Marxism: "The truth is a lie told by the ruling class; so there is no truth." Cultural Relativism: "Truth is whatever a given culture believes it to be; so there is no truth."

Setting aside the fact that any doctrine to the effect that there is no truth is self-defeating, any such doctrine is maximally easy to absorb, since, if that doctrine is right, there is *ipso facto* nothing to absorb. And to the extent that a given discipline accepts such a doctrine, its methods are *ipso facto* purely exhortatory; for, if there is no truth, then the traditional way of finding the truth---namely, objectively considering the relevant evidence and making inferences from that evidence that are licensed by rules of inference whose truth-conduciveness has been established---belongs in the ash-can.

4.3 The Psychopath's psychological architecture fundamentally different from that of the non-psychopath

In any case, psychopaths are not at all like Hannibal Lecter. Psychopaths are frauds. They can't accomplish anything without cheating. Hannibal Lecter is evil. But he isn't a fraud; he isn't a cheat; his intellectual accomplishments are legitimate.

Also, Hannibal Lecter was a *person*, in Frankfurt's sense of the word. That by itself means that he's not a psychopath. Psychopaths are pseudo-people. Psychopaths have minds, but they don't have selves. They have narratives that replace the selves they lack. This is why (intellectually minded) psychopaths are drawn towards nihilism (*nothing means anything*) and relativism (*there's no right or wrong; no person's viewpoint is any more accurate than anyone else's; no culture's values are any more reasonable than any other culture's*). As far as a psychopath can tell, *there is* no difference between truth and propaganda. So, after a fashion, he really *does* believe that all opinions *are* equally correct. An opinion is a view as to what the truth is. So, by virtue of having an opinion, one is assuming that there is such a thing as truth. That assumption is false, according to the psychopath: all opinions 100% wrong, he believes, and are therefore *equally* wrong.

The word "value" means nothing to the psychopath. He can acquire a rudimentary understanding of how to *use* value-related words. He can figure out that theft is "bad" and that saving a drowning child is "good." But he doesn't know what these terms mean. (Wittgenstein urged acceptance of the thesis that there is nothing to a word's meaning that isn't known by knowing how that word is used. (See Chapter 1 for an examination of this thesis.) In Section 6.4, it will become clear why this is relevant.) So he really *does* believe that all value-systems are equal: they are all, he believes, 100% wrong and they are therefore *equally* wrong.

4.4 Psychopaths are legion

Psychologists say that psychopaths make up around 3% of the population. There are two reasons why I put little stock in that statistic. First, it's unclear what it means. Any degradation of psychological architecture due to a loss of truthfulness with oneself is *ipso facto* psychopathy. One's psychological integrity can be intact in some respects but not in others. Also, psychological integrity comes in degrees. So, unless we're told in what respect a person's psychological integrity must be damaged, and to what degree, we're not being told anything in being told psychopaths make up 3% of the population.

Second, the diagnostic techniques employed by psychology presuppose that there are very rigid and simple connections between a person's psychological condition and his outward behavior. This, I suspect, is a by-product of psychology's attempt over the last 50 years to scientize itself. That attempt is driven by a combination of anti-mentalism and strict empiricism. These doctrines jointly constitute a framework that recognizes no diagnostic methods that don't link psychological conditions with forms of behavior so narrowly defined that instances of them are as unmistakable as the color of litmus-strip.

It follows that such tests are of limited probative value, given that different conditions can assume the same outward form and that a given condition can assume different outward forms. Also, such tests being what they are, it is easy for someone to figure out how must act, when being administered such a test, if the resulting diagnosis is to be exactly what he wants it to be. Being the con-artists that they are, psychopaths are inevitably quick to acquire precisely this knowledge.

In any case, unlike the psychologists who came up with the previously-mentioned statistic, Cleckley succeeds in giving us some understanding of the extent to which psychopathy is a part of mankind's lot:

> It becomes difficult to imagine how much of the sham and hollowness which cynical commentators have immemorially pointed out in life may come from contact in serious issues with persons affected in some degree by the disorder [namely psychopathy] we are trying to describe. The fake poet who really feels little; the painter who, despite his loftiness, had his eye chiefly on the lucrative fad of his day; the fashionable clergyman who, despite his burning eloquence or his lively

castigation of the devil, is primarily concerned with his advancement; the flirt who can readily awaken love but cannot feel love or recognize its absence; parents who, despite smooth convictions that they have only the child's welfare at heart, actually reject him except as it suits their own petty or selfish aims: all these types, so familiar in literature and in anybody's experience, may be as they are because of a slight affliction with the personality disorder now under discussion. I believe probable that many persons outwardly imposing yet actually of insignificant import really are so affected.[479]

5.0 Sociopathy vs. Psychopathy

A psychopath: somebody with no conscience.

A sociopath: somebody who has a very corrupt conscience or no conscience but who, during his formative years, *did* have a non-corrupt conscience and whose psychological architecture, developing as it did during those years, is that of a non-psychopathic person.

The mechanisms of self-censure and self-correction that distinguish the non-psychopath from the psychopath are present in the sociopath. And, in most cases of sociopathy, those mechanisms are laid down and developed before sociopathization occurs; so they're fairly hardy. Thus, at least in its structural respects, the sociopath's psychological integration is, at least during the early phases of sociopathy, more similar to the non-psychopath's than to the psychopath's.

To be sure, the sociopath's mind lacks the values that give integrity to a non-psychopath's mind. (I'm referring to psychological, not moral, integrity.) But since those values were *once* present in the sociopath, and since his psychological architecture was formed when he still had values and wasn't yet a sociopath, that architecture is basically that of a non-psychopath.

That said, it is entirely by virtue of the force of inertia that the sociopath's psychological architecture doesn't collapse. The sociopath's psychological integrity, which is what distinguishes him from the psychopath, is, at this point, purely vestigial. Consequently, there is little keeping the sociopath from becoming a psychopath.

A psychopath is a pseudo-person. He constructs a narrative that acts as a surrogate for the life-history that, had he actually had a self, he would have had. To be sure, the psychopath has experiences, just like everyone else, and he remembers them as well the next person. But he doesn't have a life's story. Much of what a normal person refers to as his "life-experience" or his "story" consists, not of his experiences, but of his reactions to his experiences: the distress they caused him; the self-criticism they prompted him; the subsequent changing of his values and his objectives; the elation he felt at having triumphed over the deficiencies in himself that led to his previous failures. These *meta*-experiences that unify a person's various first-order experiences into a cohesive structure. That structure is his "life's story." The psychopath experiences doesn't have such meta-experiences. His life is a mere sequence of experience. The psychopath values nothing. Therefore, he doesn't analyze his experiences,

unless there is some immediate practical reason to do so. He doesn't experience the internal struggles make us who we are ("who am I?," " what do I really want?," "am I going to law school because I really want to be a lawyer, or because that's the 'right' thing to do?", etc.). The psychopath is never conflicted (except in some pragmatic sense---he may not know whether or not he has to take a right turn if he's to make it to work on time). He never wonders whether he's being true to himself. For a person who values nothing, and therefore doesn't value truth, doesn't care whether he's being true himself; he doesn't even know what it is to be true to oneself. Nor does the psychopath come into conflict with other people. Of course, other people inevitably stand in his way. But the psychopath merely sees people as threats or nuisances; he is no more in conflict with them than you are in conflict with a bobcat that you're defending yourself against. The bobcat is a threat to your survival. But it isn't by representing an indictment of your psychological integration, or the values constitutive thereof, that it's a threat to you. So your altercation with the bobcat, though a serious and potentially life-altering or life-terminating matter, isn't a *drama.* There's nothing distinctively human about it. There's drama only where there is a clash of values, and there's a life story only where there's drama---- only, that is to say, where one has value-driven experiences in response to one's experiences and where, consequently, one's current experiences, thoughts, and decisions are informed, not merely by what has happened to one, but the value-driven changes that one made to oneself in response to those experiences. Since the psychopath has no values, his current experiences, etc. are not thus informed; he lives in the moment, in a pseudo-Zen manner. He has no life's story. There is no "him", really. There is a mind; but the meta-mental superstructure that is a person's *self* is quite absent.

The psychopath has to *seem* like an actual person. Otherwise, he'll be seen for the freak that he is and won't be able to gain people's trust or therefore to predate upon them. So, because he has to fit in and therefore seem human, he has to *create* a life's story; he has to manufacture a narrative. He is likely to have different narratives, since a narrative that works with one person may not work with another. Those narratives are not based on a first-hand understanding of what personhood really is. They are based on what, in the psychopath's judgment, is most likely to *work.* What sort of narrative he concocts depends on who he's dealing with and what he wants. Inevitably, his narratives are not completely false; in fact, they may be largely accurate. But that isn't because he values the truth; it's because a *pure* lie is much less effective than one that has a

component of truth; and it’s also because, if his narratives were pure lies, he would have a hard time keeping them straight and would seen be seen for what he is.

The psychopath’s self is just a script, albeit one that, like many scripts, isn’t *pure* fiction. The psychopath isn’t speaking from the heart; he’s reading from a teleprompter. Since *he* doesn’t exist, except in the same sense as Elmer Fudd, there is nothing in himself to which he wishes to be true and thus no off-message point that he wants to make. So the psychopath is unusually likely to say just what will make the boss promote him, just what will make Sally fall in love with him, just what will make Bill hand over his life’s savings to him. He’s unusually likely to be seen as “a great guy.”

One corollary of the fact that the psychopath has no self deserves special mention. When you refer to a *non*-psychopath by his name---when you refer to him as "Smith", or whatnot----the way in which you are using that expression is different from the way in which you are using that expression when referring to a pet parakeet of the same name. The parakeet is an organism that has a mind but not a self. Since personal names (e.g. "Smith"), as opposed to impersonal names (e.g. "Paris"), are, by virtue of their semantics, to be given only to beings that have selves, parakeets cannot have personal names, and attempts to give them such names, when not ironic, embody some seriously misguided views about parakeet-psychology (or about semantics). When we refer to a psychopath by his legal name, we are using that expression in the same, improper way in which we would use it to refer to a parakeet.

5.0.1 Cleckley on the nature of psychopathy

Two quotations from Cleckley will reinforce and clarify what we've been saying:

> The [psychopath's] out functional aspect masks or disguises something quite different within, concealing behind a perfect mimicry of normal emotional, fine intelligence, and social responsibility a grossly disabled and irresponsible personality. *Cleckley (1964), p. 420.*

> Attempts to interpret the [psychopath's] disorder do not, of course, furnish evidence that he has a disorder or that it is serious. For reliable evidence of this we must examine his behavior. Only here, not in psychopathologic formulations, can we apply our judgment to what is objective and demonstrable. Functionally and structurally all is intact on the outside. Good function (healthy reactivity) will be demonstrated in all theoretical truths. Judgment as well as sound reasoning will appear at verbal levels. Ethical as well as practical considerations will be recognized in the abstract. A brilliant mimicry or sound, social reactions will occur in every test except the test of life itself. In the psychopath, we confront a personality neither broken nor outwardly distorted but of a substance that lacks ingredients without which normal function in major life issues is impossible. *Cleckley (1964), p. 422.*

5.1 Rationalization: The gateway to sociopathy

Smith is a talented novelist, and he considers novel writing to be an extremely valuable activity and he also rightly considers it to be the most valuable activity of which he personally is capable. For a while, Smith tries to write novels *and* to indulge his craving for pot and junk food. But he soon finds that his pot-induced lethargy put him in an either/or situation. *Either* smoke pot *or* write novels. But don't do both. It isn't possible.

Being lazy and self-indulgent, Smith chooses to smoke pot and eat junk foot. At first, he feels terrible guilt as a result. But his guilt, though intense, isn't intense enough to overpower the forces, or lack thereof, within him that are responsible for his ditching his career as a novelist in favor of a non-career as a binge-eating junky. Still, the ever present, gnawing feeling of guilt is souring his otherwise unalloyedly blissful existence.

He thus has a choice. Either:

> (1) Start writing novels and stop smoking pot/eating junk food;

or:

> (2) Eliminate the part of conscience that makes him feel guilty for taking the easy path (smoking pot, etc.), as opposed to the difficult, but right, path (writing novels).

Taking path (2) involves Smith's *telling himself* that, in fact, he's doing the right thing, as opposed to the easy thing. It involves Smith's reconstituting his values so as to validate the weakness-driven choice that he has made. Taking path (2) involves Smith's *rationalizing*. It involves his constructing a spurious narrative relative to which the wrong choice is the right one. Such a narrative might be along the following lines[480]:

> (SR) I like writing novels, and I'm good at it. But it's morally wrong to write novels, since people's use novels to escape from ugly truths instead of dealing with them. The apparent counterexamples to this are actually the best illustrations of it. *X*, let us suppose, is a latter day Steinbeck. He writes about the insidious

> machinations of the wealthy and their devastating effects on the poor. His novels are masterpieces. But they do anything in the way of helping bring about the needed social changes? No. They do the opposite, in fact. *X*'s novels turn the sufferings of society's victims into a source of amusement. Token concern for their plight becomes all the rage; it becomes chic to wear T-shirts emblazoned with anti-establishment slogans. The corporation that manufactures those T-shirts uses over-seas sweatshop-labor. It dumps toxic waste into reservoirs. Its board members vivisect poor people for fund. So it is itself a pillar of the iniquitous social order that it was *X*'s novels describe.
>
> So, even though *X*'s heart was in the right place, he his hands are covered in blood. But his hands would be clean if he'd smoked pot, instead of writing his novels. Thus, it is my ethical duty *not* to write novels and to smoke pot all the time.

Even if SR is correct, Smith is *still* living a lie and, as a result, he is, or will in due course become, a hollow shell. For, even if SR is correct, it isn't Smith's acceptance of SR that leads him to take Path (2): it is his laziness that does so. In deferring to his laziness, Smith is alienating himself from his core values. And, in doing this, Smith is alienating himself from himself and from the sources of energy on which, when attempting to perform any sort of action, he must draw. For, as we've established, the self *is* a set of values---or, more precisely, an agency that oversees and manages an otherwise desultory stream of mental activity, giving it cohesiveness and thereby endowing that mind-constituting, but self-less stream with an enduring identity; and, as we've also established, it is through its imposition of *values* on this empty, identity-less stream of mentation that this supervisory agency accomplishes these feats. So, in so far as his conduct is rationalization-driven, it isn't self-driven, a consequence being that, so far as Smith's rationalizations play an increasing role in Smith's life-path, Smith is regressing into a condition of not having a self---of being an empty shell, who, lacking in any real identity, has to fall back on the pseudo-identities of which his rationalizations and false narratives avail him.

5.1.1 Rationalization: The gateway to sociopathy (continued)

Rationalizations tend to involve putting sincerely held beliefs to insincere use. Individual facts are not denied; but they are taken out of context, and the patterns they form are conveniently overlooked. Part of what makes it relatively easy to overlook such patterns is that, whereas one literally *sees* (or otherwise sense-perceives) raw data, patterns aren't in any literally sense seen. To "see" a pattern is to become aware of a commonality holding among many different events, between any two of which there is a gap in time and which, taken collectively, may occupy a long period of time. Thus, seeing a pattern involves making correct judgments on the basis of information stored in memories.

It follows that patterns are easily overlooked. Memories are easily corrupted, especially when they relate to emotionally touchy subject-matter. Also, the making of any judgment of any sophistication involves acceptance of an intricate network of inference rules and general principles. Given a person who is incapable of rigorous thought and is just slightly biased, or who *is* capable of rigorous thought but is deeply biased, it doesn't take much to make that person doubt the accuracy of the aforementioned constellation of beliefs. A rigorous thinker is *ipso facto* likely to have the wherewithal to produce arguments that, although hollow and sophistical, initially appear to meet his own high standards of proof and that consequently suffice to give him reason to doubt the accuracy of a once cherished belief. An unrigorous thinker *ipso facto* needs little argumentative inducement to part with a once cherished belief and therefore probably does have the modicum of argumentative sophistication needed to give himself that inducement.

Actually, I doubt that *any* real acumen is needed to rationalize; and I suspect that, although intelligent people are probably better rationalizers *ceteris paribus* than unintelligent people, the degree to which the former surpass the latter as rationalizers isn't even close to being commensurate to the degree to which the former surpass the latter as thinkers. These points are discussed and defended in the next section; so I'll hold off on them for now.

The rationalizer does not do what he thinks is right. Moreover, it's not through simple ignorance that he does the wrong thing. He's not in the same category as somebody who mistakenly thinks that smoking is healthy and therefore smokes. He's in the same category as a smoker who knows that smoking is unhealthy and tries to stave off the subsequent wrath of his conscience by concocting some narrative relative to which his hands are clean

(e.g., “It’s not as though I’m an axe-murderer---in the big scheme of things, smoking is a harmless vice,” “we all die anyway---why not do what I enjoy while I’m still alive?,” etc.).

5.1.2 Why do people rationalize?[481]

Rationalizers see the world as they want it to be, not as it is. Thinkers see the world as it is not as they want to be. So people have an incentive both to rationalize and to not think.

Also, rationalizing is the path of least resistance. Working for what you want is hard; pretending that you have it is easy. This obviously gives people another incentive both to rationalize and to not think.

Also, thinking does, whereas rationalization does not, involve recognizing that one is fallible and, therefore, that the world was not designed to validate one's opinions. Thinking involves subordinating one's mental activities to objective norms. This involves evaluating one's own thoughts and, oftentimes, admitting that they're wrong. Rationalization involves *not* subordinating one's thoughts to objective norms and *not* admitting that one is wrong. So the rationalizer's ego is spared the hard knocks that the thinker's ego has to endure.

Rationalization is a manifestation of our instinct to flee from danger. Through physical flight, we flee from an external threat; through rationalization, we flee from an internal threat.[482] Thus, a person's tendency to rationalize has deep instinctual roots. For this reason, rationalizing doesn't involve exercising one's agency. In fact, it involves letting sub-agential, instinctual forces triumph over one's agency. So far as one's agency *is* involved in the construction of a rationalization, it is nothing more than the puppet of sub-agential instincts, its sole function being to make instinct-driven, lapses in self-discipline and rationality appear to oneself to be expressions of rationality and self-control. Of course, everyone rationalizes, even the greatest of thinkers. But *to the extent* that one’s mental activity mediates rationalization, that activity is sub-agential; and *to the extent* that one’s mental activity mediates genuine thought, it *is* agential. These points are conjectural, it must be emphasized. Nonetheless, I wish to end this section with a point that is even *more* conjectural than those just made. We all know people who are extraordinarily “street smart”, even though their discursive

abilities are those of somebody whose level of intelligence is extremely low. Speaking personally, I used to believe that a street smart person was an intelligentwho didn't have the good fortune to receive a decent education; I used to believe that, *if* a given street smart person had been properly educated, he would be as discursively intelligent as, in actuality, he was street smart.

Having had a number of illuminating exchanges with a talented student of mine, Jared Murphy, I no longer hold this view. Mr. Murphy isn't a criminal in any sense of the word. But he grew up on the streets. Most of his childhood friends became career gang-members. Many of them are either dead or in jail and those who aren't soon will be. I apprised Mr. Murphy of my view that street intelligence was merely a different manifestation of the cognitive faculties that, in an educated person, are manifested as discursive intelligence. Mr. Murphy disagreed. Here is what he said *verbatim*:

> First of all, most criminals are extraordinarily "street smart", and very few non-criminals are particularly street smart. I have had many friends who chose a life of crime and who, quite early in life, become phenomenally street smart. Most of those people were not only *not* intellectually gifted, but were of below average intelligence.
>
> My experiences have made it clear to me that so-called street smart people aren't smart at all. They are not natively intelligent people who weren't educated; and had they been educated, they would have been intellectually mediocre, at best.
>
> Of course, *some* criminals are smart; the ones at the top of the heap---the Pablo Escobars, the Al Capones---are obviously highly intelligent. (I've dealt only with lower-echelon criminals; I've never met a crime-lord. But given what such people accomplish, one doesn't have to met them to know that they're smart.)
>
> But, apart from a few kingpins, criminals are, pretty much invariably, not intelligent. In fact, a prerequisite for street smarts is an absence of intelligence. The intelligent person's cognitive search space is too large. When he has to make a speedy decision, he has to rummage through so many

possible scenarios, that, before he’s made up his mind, he’s dead.

Surviving on the streets isn’t about being smart; it’s about having very highly developed instincts. The only way to have such highly developed instincts is to be an instinct-driven person. And the only way to be an instinct-driven person is to *lack* intelligence.

There are a couple of reasons for this. First, intelligent people reflect on their decisions and therefore tend to devalue their instincts. Second, for the intelligent person, thought becomes an end unto itself and the structures that mediate thought end up crowding out those that mediate instinctive responses.Thus, the instincts of a thinker cannot possibly be as finely honed as those of a non-thinker.

To be sure, I’ve known people who are reasonably smart and have reasonably good instincts. But, on the streets, “reasonably good” isn’t good enough. The margins are too thin. Things happen to quickly. And the things that happen are irreversible. On the streets, you aren’t given two weeks' notice; you’re given two in the back of the head. No matter how little formal education a person has had, that person is a dead mean on the streets if his natural bent to is to be a thinker. That person will start to question his lifestyle. As a result, he'll be less "in the zone" than his unintelligent counterpart and therefore less attentive to the nuances to which, as a criminal, he must be attentive and less quick on the draw in situations where the first person to pull the trigger is the one who lives. So, long story short, street smart people aren’t smart. They can't be. They have to be people whose only option is to hone their animal instincts. If they have any intelligence, that won't be their only option.

Since, unlike Mr. Murphy, I didn't grow up on the streets, I don't whether he's right or wrong. But my suspicion is that, while there is almost certainly an element of hyperbole in his words, there is more truth than falsehood in them.

5.2. Why rationalization is inevitable and often healthy

Everybody rationalizes. Setting aside people who are exceedingly damaged (e.g., people who have sustained serious brain-damage) it is psychologically impossible for a person not to rationalize.

Moreover, it isn't always bad to rationalize. In fact, it would be bad to *never* rationalize. There would be decision-paralysis if the brain-centers tasked with making decisions were flooded with *too* much truth. Also, one's conscious mind can only withstand a modicum of truth, lest one's self-concept sustain irreparable damage. Speaking from an evolutionary-biological perspective, knowledge is a means to an end, the end being health. A complete absence of conscious knowledge isn't health-conducive, but neither is an excess of such knowledge. One has to make decisions in real-time. One's "search-space" can't be too large. Moreover, so many truths of which one is aware, at some level or other, are so ugly that one cannot confront them all at once and must do so in a piecemeal fashion.

People can't deal with everything at once. So they compartmentalize. Sometimes bodies of information that are conscious (i.e., are readily accessible to consciousness, not being held down by the forces of repression) are compartmentalized. But this form of compartmentalization is unstable. Two pieces of information both of which one is consciously aware, tend to coalesce. There's only so long one can consciously know both (a) that one's good buddy Smith is the only person who could have put arsenic in one's tea and (b) that one has been hospitalized repeatedly for arsenic poisoning without also knowing (c) that Smith isn't really one's good buddy.

Also, there isn't a sharp line between rationalization, on the one hand, and the legitimate employment of rationality on the other. There are three reasons for this:

> (1) The construction of rationalizations tends to *involve* the legitimate employment of rationality.
>
> (2) As previously stated, some rationalizations are *true.* And, in at least some cases, the difference between a belief that results from rationalization-based belief and a 'legitimate' belief isn't that the former is, whereas the latter is not, true; nor, necessarily, is it even that the former is, whereas the latter is not, arrived at in

a spurious manner. In at least some cases, the difference lies in the *use* to which the belief---which may be true and legitimately arrived at—is put. And a given (legitimate and legitimately arrived at) belief can be put to different uses. It can be used to figure out how to help struggling nations become more prosperous and it can also be used to figure out how to impoverish those same nations.

(3) A given belief can be a rationalization in one respect and a non-rationalization in others. One may have different reasons for arriving at a given belief, some of them embodying wishful thinking and others embodying sober consideration of the facts; and one may use a given belief as the basis for a rationalization and for a legitimate pursuit of the truth. You believe that you mom is a good person partly because you want to believe that and partly because it's true to a large extent, though not as large an extent as you'd like to think, and the evidence supports it, though not as much you'd like. And you can put your belief that your mother is a good person to different uses: you can use it as an excuse to look the other way when she does something that a good person would not do (rationalization) or you can use it to subject a claim made about her to sober examination (non-rationalization); e.g. on the basis of your belief that your mother is a good person, you may accept the rationalization that her not allowing you to do athletics wasn't intended to emasculate you or to vent her rage at your father (who was a professional athlete and incessantly carried on with other women before dumping your mother); and on the basis of that same belief, you can determine that so and so is demented or deliberately untruthful, since he routinely makes claims about your mother (e.g. "she eats little children") that plainly aren't true.

5.3 When rationalization is not healthy

Given that everybody is, to some degree or other, a rationalizer, what is the difference between those rationalizers who sociopathize and those who don't?

There are two differences:

> (1) For non-sociopaths, rationalizations are misfortunes, not deeds. For sociopaths, they're deeds, not misfortunes. Non-sociopaths don't seek out rationalizations; rationalizations are sought out *in them* but not *by them.* In the non-sociopath, rationalization happens sub-agentially. In the sociopath, rationalizations are deeds; he seeks them out, perfects them, and accepts them. In the sociopath, rationalizations *also* occur sub-agentially. So the difference between the sociopath and the non-sociopath isn't that the faculties responsible for the construction of rationalizations have been *transferred*; it's that, without losing any of their original (sub-agential) jurisdiction, they *spread*: they appropriate the subject's agency, leading in some cases to a complete reconstitution of the self.[483]
>
> (2) Non-sociopaths care about the truth. Sociopaths do not. The non-sociopath's concern for the truth keeps his personality intact; it insulates it from his rationalizations. The sociopath's self has no such insulation. As a result, it is eventually taken over by its own rationalizations.
>
> The sociopath, unlike the psychopath, was once a person. The sociopath's personality-architecture, at least during the initial stages of sociopathization, is far more like that of non-sociopathic non-psychopathic than is the personality architecture, what rudiments of it there happen to be, of the psychopath. When the soon-to-be sociopath stops caring about the truth, the rationalizations that swoop in occupy the same positions in that personality as the beliefs and values that they displaced. Thus, for a while at least, a sociopath's rationalizations operate within the same sort of structure as a non-sociopath's genuinely held beliefs. But that structure embodies values that the person in question no longer has, and it eventually collapses.

(2) is a corollary of (1). If a given person has no concern for the truth, there is nothing preventing him from actively seeking out rationalizations.

5.3 Sociopaths incessantly moralize

Sociopaths are very moralizing. When speaking with them, one cannot be oneself; one must behave in a hyper-correct fashion, lest one incur their virulently ill-willed, pseudo-moral vituperations Being with a sociopath is like being in *Oceania*, the hyper-totalitarian society described in Orwell's novel *1984*. Just as, in *Oceania*, one is vaporized if one has any resistance to accepting the government's lies, so, when speaking with a sociopath, one is censured, and thus vaporized, as it were, if one doesn't operate within the sociopath's candy-coated, morality-drenched narrative.

But the morality is fake; it has nothing to do with the sociopath's beliefs as to how the world ought to be. This fake morality has four sources.

First, sociopaths don't want people to see them for the vampires that they are. They want people to think that they're good. It isn't only for strategic reasons that they want this. The sociopath has nothing but his image. Whatever substance he one once had has been drained away and replaced with a narrative. He's reading from a teleprompter, not speaking from the heart. There is no heart left. There is no self, in other words. So sociopaths need the approval of others to feel that they have selves. And that's why they insist on saying "correct" things, instead of human things. It's also why they're drawn to "glamour." Show business is replete with sociopaths. Second, because sociopaths are so alienated from what is (or, when it existed, was) human in themselves, they are forced to "phone it in": they have to read scripts; there is no longer a person, a self, within them for whom they can speak; if their statements expressed their true inner lives, it would consist of nothing but the tirades of an infant who hasn't yet acquired a self and whose emotional life is driven by infantile narcissism and, for that reason, sheer detestation of those who interfere with their infantile aspirations.

Third, sociopaths resent their condition; they resent selling their souls; they resent being made to give up their selves; and, for that reason, they're envious of those who have hung on to their selves. So they lash out at those people---but, not wanting to be seen for the empty, tyrannical frauds that they are, they do so under the banner of morality. Which, if the recipient of their morality-coated rage is naïve, will hurt him and which, at the same time, will help nobody, since its only real effect is to send the message that it's better to be a non-entity who wears a diamond-studded mask (underneath which there is no self, only a disorganized mixture of self-hatred, compensatory rage, and

infantile (and also shame-driven) feelings of entitlement) than it is to be the real person who is who he is.

Fourth, for the sociopath, morality is a weapon. By being moralizing, sociopaths have what they've always wanted but never had: power. Powerlessness makes one sociopathize; and once one has sociopathized, and has thus overthrown one's conscience, one has no compunctions about using morality to belittle others.

Earlier we mentioned Orwell's book *1984.* Speaking from a psychodynamic viewpoint, *1984* is not fiction. It is an accurate description of mass sociopathization. If the forces at work within the real world were embodied in our political institutions, the result would be the world that Orwell describes. *1984* is not primarily about politics; it's about people being driven by fear into believing what they know to be lies and, as a result, ceasing to be human. Fortunately, a system of government in which there are checks and balances is capable of inhibiting, and sometimes altogether preventing, the sociopathization of the political institutions on which people depend for their welfare.

5.4 Vampirism a metaphor for sociopathy

Many movies and books are about vampires. Vampires are typically represented as having certain properties.

(i) Vampires don't appear in reflective surfaces, such as mirrors.

(ii) Vampires are creatures of darkness. Sunlight kills them.

(iii) A vampire cannot see his own face. (This follows from (i).)

(iv) Only others can see a vampire's face (This follows from (iii).)

(v) Vampires live in opulence; they are fabulously wealthy. Their spacious domiciles are richly appointed. They have all the material trappings of success. They are elegant and well-cultured. They dress like royalty.

(vi) Vampires can turn into bats.

(vii) One becomes a vampire by selling one's soul.

(viii) A vampire must drink the blood of the living; otherwise he dies.

(ix) Vampires are parasites and thus have only a borrowed, second-class existence. (This follows from (viii).)

.

Taken jointly, (i)-(ix) suggest that vampires represents sociopaths.

Interpretation of (i) and (ii): Vampires have no substance. That's why they cease to exist in the light of day, and it's why they don't appear in mirrors. Like vampires, sociopaths have no substance. That's why, when scrutinized---when looked at in the light of day, as it were, and not allowed to take refuge in darkness--- the sociopath's personality vanishes, his morality and indeed his entire personality proving to be nothing more than fictions.

Interpretation of (iii): The vampire's face represents the sociopath's self-concept, and the vampire's inability to see his own face represents the fact that

the sociopath has no endogenous self-concept. Lacking in any real self-concept, sociopaths are empty; there is no life in them.

Interpretation of (iv): The fact that only others can see the vampire's face corresponds to the fact that the sociopath's self-concept is dependent on the acknowledgement of others. (Sociopaths tend to spent a lot of time in front looking at themselves in the mirror.)

Interpretation of (v): The sociopath is empty; he tries to fill the void with external tokens of the success. Also, since can't distinguish the image from the reality, and wealth creates the image of success, the sociopath equates wealth with success.

Interpretation of (vi): Bats are blind. So are sociopaths. Sociopaths believe what they want to believe, not what is real; they are blind to the truth.

Interpretation of (vii): One becomes a sociopath by selling out.

Interpretation of (viii) and *(ix):* The sociopath's existence is a waking death, from which he has no respite except when he's drinking up the adulation of others, a corollary being that, without the adulation, the sociopath suffers the same fate as the blood-deprived vampire.

5.5 The professor as sociopath

I will try to answer this question by answering a more specific one, namely: what kind of a person is a sociopathic *professor* likely to be. Why take this approach? First, since I work on a daily basis with many sociopathic professors, and since, being a professor myself, I am inevitably made privy to their machinations, I am able to speak authoritatively about them. Second, because my sociopathic colleagues are strikingly similar to the many sociopathic non-professors with whom I acquainted, even though the latter come from all walks of life and are engaged all manner of professions, I have reason to suspect that those character-traits are inherent in sociopathy itself and, therefore, that points similar to those that I will now proceed to make about professors hold, in some form or other, with respect to sociopaths in general.

Sociopaths are weaklings who can't hack it and then lie, first to others and then to themselves, in order to cover up their deficits. They are people who aren't great but are so afflicted by infantile narcissism that they cannot tolerate *not* being great. People of integrity who'd like to be an Einstein or a Beethoven but find out that they're not in that league don't become sociopaths; they become people who do what they *can* do and they take pains to do it properly. The non-sociopath who wishes he were Einstein, but knows that he isn't, becomes a capable and hard-working civil engineer. The non-sociopath who wishes he were Beethoven, but knows that he isn't, becomes a capable and hard-working film-composer or music-teacher. The sociopath chooses to be a fake Beethoven rather than a real sub-Beethoven. He chooses to be John Cage, whose work is bogus but chic, rather than Paul Simon, whose work, though sub-Beethovenian, is superb.

Sociopaths claim to be rebels who "think outside the box." But they're not rebels; in fact, they're arch-conformists. And they don't think outside the box; in fact, they don't think at all. They embrace positions (e.g. Marxism, Cultural Relativism, Post-Modernism) that, in some cases, were once rebellious but, at least in the contexts in which the sociopath advocates them, are not only *not* rebellious but shield their advocates from the scrutiny and criticism to which their opponents are exposed.

If a professor takes the post-modernist view that truth is culturally relative, he'll be applauded; if he takes the view that truth is not culturally relative and that, in some cases, entire cultures have the wrong view, he'll be derided by his

students and censured by his superiors. If a professor takes the radically implausible post-modernist view that the psychological differences between men and women have no biological basis, he'll be praised for his open-mindedness; if he takes the opposite view, he'll be branded a moron and chauvinist pig. If a professor takes the Marxist view that so-called moral verities are but artifices by which the haves manipulate the have-nots, he'll be seen as a cutting edge thinker who is willing to question "dominant paradigm." If a professor takes the non-Marxist view that moral truths are not mere chicanery and that they reflect opinion-independent facts about what conduces to the health of individuals and groups thereof, he'll be seen as a pointy-headed, retrograde square. Also, if one takes the post-modernist view that there is no truth and, more generally, that none of the benchmarks in terms of which human beings evaluate themselves have any legitimacy, one thereby insulates oneself from criticism; for any criticism of his views presupposes that there is truth and is therefore spurious.

The chic Marxist law professor who, in the U.S. in the year 2010, advocates the downfall of the U.S. government isn't a rebel at all. He knows that his words will have no effect at all, except to misinform his students and stock the empty heads of faux-intellectuals with ready-made slogans that they can use at cocktail parties, and he knows that the very government whose downfall he advocates will ensure that not one hair on his head will be harmed. Also, unless he's so totally sociopathized that he himself is completely convinced by his own mythology, he knows that his proposals, if by some miracle put into action, would obliterate the social structures on which life as we know it depends. So he's a fool if he believes what he's saying and a fraud if he doesn't. Finally, given the almost certainly correct supposition that he knows, at least some degree, that what he's saying isn't true, he knows that he's misinforming his students, weakening their minds, and undermining their autonomy.

Similar points hold with respect to the hip media studies professor who says that only suckers read books by dead, white males and that, were it not for the machinations of advertising executives, we'd realize that listening to hip-hop was better for our intellects than reading John Stuart Mill. This enlightened hipster is saying that logic and hard-work mean nothing. This message may sound liberating. It is not. It's a way of telling people to cripple themselves. ("Deprogram yourself. Whatever part of you is responsible for what you naively thought of as thinking—rip it out!") Thus, it's a way of telling them to cease to exist. For nothing remains of a person once he ceases to be able to think for

himself.

Like the Marxist law professor, the media studies professor hobbles his students under the banner of emancipating them. The same is true of the women's studies professor who contends that the psychological differences between men and women are entirely the result of social conditioning. This contention entails that, because women have been brainwashed, they can't think for themselves and, therefore, that anything they have to say is to be disregarded. It also entails that, since the dawn of civilization, men have succeeded in duping, not just a few women, but *all* of them, with a tiny number of exceptions. This in its turn entails that women are, by nature, weak and suggestible. It also entails that men seriously outclass women intellectually. And this, in *its* turn, undermines the contention that, were it not for the lies told by men, women would be just like men. For, were this thesis true, it wouldn't have been so easy for men to manipulate women.

In fact, that hypothesis, when considered in isolation of the contexts in which it is advocated, has decidedly retrograde and anti-feminist undercurrents. *Men are brutes, who through charm and cunning, manipulate women into being their willing servants*. This sounds like the central conceit of a Harlequin romance, and it hardly seems a viable basis of a political movement whose objective is to emancipate women from the shackles of archaic, misogynist attitudes and the social structures that perpetuate them.

So one in saying that, were it not for misogynist propaganda, women would be psychologically just like men, one is saying, if only by implication, that

> (i) men are naturally more cunning and treacherous and women and, consequently, that men and women *are* psychologically different;

and also that

> (ii) women by nature easily taken advantage of.

"But you're reasoning is quite off the mark," it will be said. "First of all, supposing that, indeed, men have been dominating women through the ages, they may well have been doing so through sheer physical force. Second, it could just be a historical accident that men happened to be the first gender to use propaganda as a way of controlling the opposite gender; so the fact that, thanks

to their chicanery, men are "on top," as it were, may reflect nothing more than happenstance."

First of all, the contention being discussed is that it was not through force, but through *propaganda*, and therefore through trickery, that men came to dominate women. Second, given that men are the, for lack of a better word, dominant gender in *every society*, many of which were isolated from the others, shows that, if indeed it was through chicanery that men came to dominate women, it wasn't the product of historical happenstance and was instead the result of some constitutional difference between men and women.

Let us move one. Women have nothing to gain, and much to lose, by accepting the contention that, but for male propaganda, women are psychologically just like men. For a woman who accepts that view *ipso facto* believes that *she* has been brainwashed and, therefore, that she can't think for herself, and she *ipso facto* thinks the same of other females. ("But women who realize that they've been brainwashed," I've heard it said, "are *ipso facto* free of the effects the brainwashing to which they've been suggested." This is deeply implausible. So far as it's even possible to completely undo the effects of a lifetime of brainwashing, it can't be done overnight: realizing that one has been brainwashed is only the first step in a very long journey whose final destination is, at best, only a partial restoration of one's faculties.) Thus, a women who accepts this supposedly progressive thesis necessarily believes that men are the only people who have not been brainwashed and therefore that, whatever moral deficiencies men might have, they are the only people who are even capable of seeing things as they are or, therefore, of providing sound guidance. So such a woman undervalues herself, along with her fellow females, and she over-values men. This double irony has a grim, as well as ironic, consequence, namely, that a woman of the kind just described is an unusually easy mark for male predators.

It's obvious that "so and so has been brainwashed" entails "so and so's judgment has been impaired." In fact, the second statement is a tautologous consequence of the first, given that it makes no sense to say that somebody has been has been brainwashed but that his powers of judgment haven't been negatively affected. So the professors who say that, for the last several thousand years, women *en masse* have been brainwashed must surely know that, in saying this, they're saying that women are intellectually stunted. Thus, such professors cannot possibly be motivated by a wish to help the plight of women and can only be motivated by hatred of them or a by characteristically

sociopathic willingness to base one's career on a lie.

Each of the three pseudo-iconoclasts just described is driven by shame: shame at not being able to produce a legitimate line of goods; shame at being sterile and impotent. Having failed relative to the yardsticks by which it was their original intention to succeed, they embrace an ideology that invalidates those yardsticks. (Post-modernism: the standards in terms of which we evaluate what we say and do have only as much validity as we want them to have. Relative to one standard, modern medicine is superior to Neanderthal medicine; relative to a different, but equally valid standard, it's the other way around. Marxism: the truth is a fairy tale told by the haves to manipulate the have nots. You didn't really fail that test----that's just what The Man wants you to believe.) Given how brazenly false such ideologies are, their advocates cannot accept them without ceasing to value truth and thus, given how much of a psychologically healthy person's identity depends for its existence on that person's valuing the truth, without eviscerating themselves.

These fraudsters thus have multiple incentives to convince others of the truth of the falsehood to which they've sacrificed their identities.

First, misery loves company: they want others to make the same sacrifices they've made so that others will suffer with them.

Second, they desperately need others to tell them they're *not* frauds. This is because the doctrines they espouse are *so* at variance with such obvious facts that, deep down, they don't really accept those doctrines, notwithstanding the psychosurgery they've performed on themselves.

Third, they need others to adulate them. The reason for this is different from the reason for which they need others to tell them that they're not frauds. People who value the truth can validate themselves. An author who values the truth can take pride in a book he's written that nobody ever reads but that he knows to be good. As somebody who values truth, he knows that any praise that he *were to* receive for his book would have no more value (except, of course, of a purely instrumental kind) than the book itself. So, since he already knows his book to have value, any praise that he might receive for it would (except from a practical point of view) be redundant.

But, to somebody who doesn't value the truth, true statements are no more valuable than false one. A corollary is that, as far as such a person is concerned, it's irrelevant whether the praise they receive is well founded. Well-founded praise is truthful praise. Since sociopaths have to believe their own spurious narratives, they identify the appearance of truth with truth itself. A writer whose

works aren't read *appears* to be a failure. So, from the sociopath's viewpoint, he *is* a failure. A writer whose works are best-sellers *appears* to be a success. So, from the sociopath's viewpoint, he *is* a success. Since the image is the reality, a book is only as good as it is praised, and a person is as good as he is adored.

Though he doesn't want be adulated, the non-sociopathic literary genius wants to be *recognized.* Adulation is different from *recognition.* Recognition=being praise because of what one has done. Adulation=being praised, not because of what one has done, but because of who one is. For Smith to be adulated is for Smith to be praised for being Smith. For Smith to be recognized is for Smith to be praised for composing a great symphony.

People of conscience want see *justice* done, and they thus want credit to be given where credit is due. To be sure, there is an element of sheer egoism in a writer's desire for recognition. In fact, his egoism is almost certainly likely to be much more of a fact than his sense of justice. But the way in which egoism is expressed in a person of conscience is very different from the way it is a expressed in a sociopath. The writer's egoism is expressed in his wanting to be praised and in his not caring very much whether other, equally deserving writers are praised; his sense of justice is expressed in his wanting to be praised *because he deserves it.* Many writers, thinkers, and artists of genius are praised for works of theirs that they rightly regard as among their worst. For example, Rachmaninov, who was a musical genius of staggering proportions, was best known for his prelude in C# minor, which, as he knew, was one of his worst works. For financial reasons, he had to perform it. But, whenever he did so, he went into a depression. Although I have read extremely little good literature and have no eye for graphic art, I am quite certain that Rachmaninov's situation is by no means unusual. The reason that Rachmaninov found it so dispiriting that he was praised for this work of his is that he wanted to be praised for the right reasons: he wanted the commendations that he received to be ones that he earned; if only within the confines of his attitude towards people's feelings towards his work, he wanted *justice*.

Rachmaninov's attitude towards the misguided adulation that he was received was quintessentially *non*-sociopathic. The sociopathic wants praise for its own sake. He doesn't care whether he's earned it or not. In fact, he doesn't really have a concept of *earning* anything. For if one distinguishes between those who have things because they *earn* them, and those who just have them, one must distinguish between image and reality; and the essence of sociopathy is a refusal, driven by a need to believe one's own lies, to make the latter

distinction

"According to you,", it is likely to be said, "authors who know that they're work is good don't want praise. But that isn't true. And surely that discredits much, if not all, of what you've just said."

The first part of this point is quite right: many an author who values the truth wants adulation despite knowing that what he's written is good. But this doesn't undermine anything of I've said. This is because, when such an author wants praise, it is for one of three reasons:

> (i) Although he knows *at some level* that what he's written is good, some part of him skeptical; and, in being praised, his insecurities are finally being put to rest.
>
> (ii) in being praised, this previously unnoticed literary giant is finally being given evidence that there is at least some justice in the world;

or

> (iii) such praise brings with it all manner of monetary and practical benefits thanks to which one doesn't have to waste time groveling before slimy editors and working at a menial job and thanks to which one finally has access to some of the finer things in life.

But this only confirms our point that, if a person genuinely is convinced that what he's written is good, he doesn't want adulation *per se*: what he wants are the benefits that adulation brings with it.

Some are likely to reject (i). "If an author knows that what he's written is good," it will be said, "he *ipso facto* cannot possibly be at all insecure about its merits." This criticism falsely assumes that people cannot have mutually opposed beliefs. The statement "Smith believes that 1+1=2 and Smith does not believe that 1+1=2" is incoherent. But the statement "Smith believes that 1+1=2 and Smith believes that 1+1≠2" is perfectly coherent. In fact, the history of ideas is replete with cases of people having incompatible beliefs. In particular, many a great artist (e.g. Brahms) was rightly convinced, in a genuine and non-defense

way, of the excellence of his work while being tormented by terrible insecurities about it. People are not cognitively unified. And an author who knows that his work is great can, at some level, believe that his work is bad and, therefore, will find praise for his to work to be a source of comfort.

5.6 Fame as sociopathogen

Fame is a great sociopathizer. People who, within narrow limits, had merit are, for some random reason, propelled to the heights of stardom, at which point they are grossly overvalued by the public and start to feel like frauds. They then have a choice:

> (i) They can believe the lie that everyone is telling them (*you're great, you're a real thinker, not just some comedian whose sense of humor was so anodyne that some board of directors knew it risked nothing by making you a talk-show host*);

or:

> (ii) They can admit that they're not great and then either renounce the benefits of fame or consciously live a lie.

One can understand why, confronted with this choice, many people go with (i).

6.0 The non-sociopathic/non-psychopathic predator

A person who doesn't rationalize---who doesn't replace his true self with a fake narrative---is not a sociopath. Non-sociopaths are sometimes evil. Some of them, in fact, are supremely evil. In fact, if a list were made of the 10,000 most evil people of all time, not a single sociopath would appear on it. (Nor, for very different reasons, would a single psychopath be on that list.) This is because sociopaths *want* to believe that they're not as empty as they really are. And, despite their willingness to prefer fiction over truth, they'd make it very hard for themselves to believe that they had merit *if* they acted as treacherously as a Stalin or a Mao.

"Sociopath" doesn't mean "evil person." "Sociopath" means "fake person with a corrupt conscience." ("Psychopath" means "non-person who never had a conscience and who, for that reason never had a self, has always lived through narratives fabricated *ad hoc*.") Sociopaths are always evil. But evil people are not always sociopaths. And evil non-sociopaths are *more* evil than sociopaths, since, unlike sociopaths, the former aren't struggling to convince themselves that they're good and thus have no compunctions about going all out, as it were, when it comes to evil-doing.

Sociopaths aren't human. They've stopped being human. Evil non-psychopathic, non-sociopaths are thoroughly human. But, whereas my objective is to write treatises and yours, let us suppose, is to become a pediatrician, their objectives are to do harm.

6.1 The non-sociopathic/non-psychopathic predator (continued)

“How is it possible," it will be asked, "for that to be the objective of somebody who isn’t deranged in some way?”

The right question is: “How is it *not* possible?” [484]

We’re the victors of a long, brutal evolutionary struggle. We chose to fight it out. We didn’t opt out, like clams, by becoming so lame and harmless that we weren’t worth destroying; nor did we become peaceful herbivores whose entire lives are but preludes to being slaughtered and eaten. So it’s no surprise that there are at least a few predators among our ranks.

Would a tiger be better off if it renounced violence? No! Tigers must be tigers; if they try to be sheep, they’ll die on the inside. Would a born predator be better off if he were a Buddhist monk? No! Like the tiger, he’d die on the inside; and, in all likelihood, his need to predate would come out in circuitous ways (e.g., pettiness towards underlings, thwarting those whose aspirations are legitimate and whose work-product is authentic). By trying to be peaceniks, these born war-mongers either injure themselves or, by turning structures on which the public depends for its welfare into outlets for their undischarged aggression, do harm on a far greater scale than they would have had they discharged their aggressions more directly.

6.2 Criminalization as a way of avoiding sociopathization

Some people who become professional criminals (e.g., members of organized crime) do so to *avoid* sociopathizing. In their view, working for "the man" involves throwing oneself into a sausage-grinder: one must tow the company line and, to use a phrase from Orwell's *Keep the Aspidistra Flying*, "lick the blacking off the boss's boots." It involves becoming an empty suit: a hollowed out shell who believes nothing except what higher ranking suits want him to believe.

Winston Smith, the main character in *1984*, says how he hates "goodness" and "purity." What he means is that he hates the fake goodness and the fake purity with which, as a citizen of Oceania, he is inundated. What he means, in other words, is that he wants to be human, not "pure" or "good", which words, as they are used in his society and as they are used by the sociopath, *don't* refer to purity or goodness, but to their plastic, sociopathized analogues.

Winston was a good and honest person in a world of criminals. So why did he say that he hated purity and goodness? Ms. Marks' insight can help us answer this question and, in so doing, can help us understand the mentality of non-sociopaths who operate outside "the system."

Former Mafiosos who joined the Witness Protection Program don't rationalize. Sammy "the Bull" Gravano said that he was never, at any point, to any degree, a "Robin Hood" whose intentions were to do good; he was, he said, somebody who wasn't educated but who wanted power and was willing to "scratch and kick and do whatever it [took]" to acquire it. Other former Mafiosos, without few exceptions, say the same thing. Jimmy "the Weasel" Frattiano said that he was a less than stellar human being who chose a life of crime *not* because he had no other options, but because he liked having power more than he liked being moral and, he implied, he would have been a powerless nobody had he "gone legit."

People who admit to themselves that they're evil are to that extent *non*-sociopaths. The essence of sociopathy is a shame-driven rationalization of one's inability to live up to one's own values. So far as one genuinely "owns" one's actions, which oftentimes involves admitting that one is a bad human being, one is not a sociopath.

In this context, it must be stressed, we are discussing people who criminalize *in order to stave off* sociopathization; we are not discussing sociopaths who

criminalize in order to hide their condition from themselves. Many people *become* predators as a result of sociopathizing; and they falsely tell themselves that, in going this route, they were simply doing what the laws of the jungle required them to do. ("It's a rough world. Nobody does you any favors. Do whatever it takes. Just don't get caught.") Such people are lying by pretending to be brutally honest: their pseudo-honest admissions to themselves and others that they are predators are just ways of cloaking their weakness-driven descent into sociopathy.

But in some cases, as we've said, one criminalizes in order to hold onto one's identity, and therefore in order to avoid sociopathizing. Since sociopaths aren't who they are---since they half-believe their own propaganda and thus half-believe they *are* good---there tend to be limits to how evil they can be. But there are no limits to how evil a non-psychopathic, non-sociopath can be. For example, Joseph Stalin---who, by any measure, was one of the most wicked human beings ever---was not a sociopath. He was a serial killer who, having managed to acquire an enormous amount of power, was able, unlike other predators, to predate on a large scale and without fear of legal reprisal. He knew he was evil; and he rejoiced in it. He didn't deceive himself; he didn't say that "it was for their own good" that he displaced entire populations and condemned millions to starvation. He did it because he wanted to do harm.[485] To accomplish this, he obviously had to lie to others. But he saw himself for what he was: a mass-murderer who also happened to be a head of state. So he wasn't a sociopath.

Although Stalin *was* a national leader, he saw himself for what he was: a butcher, who, through chicanery and deceit, obtained for himself a position from which he could, with impunity, be the serial killer the he was. So Stalin was not a sociopath; for, although he lied to others (on an unprecedented scale, in fact), he didn't lie to himself (except in so far as everyone does).

Nor was Stalin a psychopath. A psychopath couldn't have done what Stalin did. Stalin's ascent to power involved his being able to evaluate himself with respect to yard-sticks that weren't concocted *ex post facto* to rationalize his errors. He wouldn't have learned from the past if, when evaluating his mistakes in terms of yardsticks that were *independent* of his desires, he weren't able to see the folly of his ways.

Also, psychopaths are extremely impulsive. They can contain this side of themselves for a while; but sooner or later it comes out. And when it does, their careers are over. Their impulsiveness is a consequence of their having only a

limited and contextual grasp of the truth. Since, as previously stated, they don't value truth for its own sake, they value it only when they *know* that it will serve them well. And since, as we observed, it's impossible to prejudge when some now useless truth will become useful, the result is that the psychopath has an excessively limited body of truths from which to draw when making decisions. Because he is thus blinkered, his impulses aren't countered by a firm grasp of their consequences. Hence his impulsiveness. Stalin was not impulsive. He was calm and rational; his decisions, though maximally evil, were judicious and calculated.

Etymologically, the morpheme "path" (as in "sociopath" and "psychopath") means "disease." Sociopathy and psychopathy are both form of pathology. That said, psychopathy is a much more acute form of pathology than sociopathy. The sociopath's psychological architecture is in the process of imploding; he's the psychological equivalent of somebody whose disease-ridden internal organs are only semi-operational but have retained their basic shapes. By contrast, the psychopath is like somebody who never developed those organs or who did develop only for them to be liquefied by a case of Ebola.

Stalin didn't fall into either of these categories. He knew what he was; he knew what he wanted. What he was: a predator; a person of some intelligence, but not as much intelligence as others; a person who, if he was to predate without being predated upon, had to engage in all manner of chicanery. Since he had a self, he wasn't a psychopath; and since he didn't tell himself that he was something that he was not, he was not a sociopath. He was just evil. [486]

6.2 Donald Davison's analysis of what intentions are

According to Donald Davidson (1963), an intention is simply one's strongest desire. More precisely, it is one's strongest contextually relevant desire. So, if I am in an ice-cream store and, for some reason, it's incumbent on me to eat ice-cream, my subsequent *intention* to eat chocolate ice-cream is the strongest intention that I have that is an intention to eat ice-cream at that time, in that place; it is stronger than my desire to eat any of the other kinds of ice-cream that are available. At the same time, of course, my desire to eat chocolate ice-cream is not my strongest desire *tout court.* (In what follows, any references to one's "strongest" desire are always abbreviated references to one's strongest contextually relevant desire.) So, according to Davidson, an intention is the desire that, because of its strength, beats out the other desires with which it is in competition and thus, unlike them, expresses itself in action.[487]

This reasonable-seeming analysis embodies several confusions. (I should point out that, in his excellent book *The Faces of Intention*, Michael Bratman makes points similar to, but not coincident with, those that I am about to make.[488]) A desire that is directly converted into overt behavior is a reaction, not an action proper. For a desire to lead to action, it must be inhibited and scrutinized, so as to ensure that it embodies the right values. By virtue of surviving such scrutiny, a desire becomes an intention and, to the extent that circumstances allow, automatically discharges itself in action.

Even those desires that are expressed in action *despite* not surviving such scrutiny are not cases of desires being directly converted into action. They are cases of desires being scrutinized, then found to fall short, but being allowed to discharge themselves anyway. The reason the tobacco-addict feels ashamed is that, despite what his conscience tells him as he's considering his wish to smoke, he nonetheless lifts the initial inhibition that he has on smoking. And even though, in respect of his overt behavior, he may pounce on the cigarette the way a tiger pounces on red meat, the addict's behavior is that of somebody in whom the impulse to smoke is the result of a *decision*: his impulse is initially quarantined, then scrutinized, and only *then*, despite its falling short of the relevant valuative benchmarks, allowed to be discharged in action. The tiger's desire, on the other hand, is directly translated into action. Where selves are concerned, few so-called impulsive behaviors, if any, involve a desire's being directly translated into action; they typically

involve cases where there is an initial inhibition which, despite his better judgment, the person in question chooses to lift.

That said, the decision to lift the inhibition sometimes *is* a wise one. The person in question wishes to *appear* like a violent, unpredictable animal, since he knows that, given the circumstances, this is to his strategic advantage; and he knows that, if he punches the person who just insulted him, that is precisely how he will appear. Many apparently impetuous actions are carefully considered. We must remember that "carefully considered" ≠ "slowly considered." A carefully considered decision can be made virtually instantaneously. This is partly because we think fast. But it's primarily because, given previous decisions that we spent a great deal of time making and that we've decided to comply with, it's immediately obvious how to behave, when contingencies arise, in a manner that is consistent with those decisions.

Our desires are fleeting. They constantly change. If Davidson were right, we'd be buoyed hither and thither by our desires. Every time a sufficiently strong desire came along, we'd "act out." But, in so doing, we wouldn't really be acting at all. For our overt behavior wouldn't be the outcome of a decision.

In fact, if Davidson's theory were correct, *we* wouldn't be involved at all. Our impulses would use us to express themselves.[489] The overt behavior in which that desire was issued wouldn't be an action at all, at least not in the relevant sense. (It would be an action only in the sense in which a tiger's instinctual reactions are actions; it wouldn't be an action in the sense in which your rational and carefully considered decision to become a physician is an action.)

To be sure, if a desire is overwhelmingly strong, we will tend to act on it. But, setting aside *bona fide*psychopaths, what tends going in such cases isn't that the desire *directly* results in action, and is instead that the screening process leading to the lifting of the relevant inhibition is fast-tracked and, possibly (but not necessarily[490]), done shoddily.

Contrary to what Davison holds, an action isn't the direct expression of *any* desire, be it one's strongest desire or one's weakest. An action is the direct expression of a value-driven decision to lift an inhibition on a desire. When deciding whether to lift that inhibition, one considers how doing so would affect one's self-concept and one's overall life-course.

Thus, an intention, contrary to what Davidson holds, isn't a desire at all. It

is a *judgment.* That judgment is intended to gratify some desire; but the desire is not the intention. To sum up: if Davidson were right, we'd never act; we'd only react. We'd be like pinballs, bouncing from bumper to bumper, each bumper being our strongest desire at the time.

But that isn't how we are. As Bratman says, we make plans. Once we make these plans, we tend to stick with them. Desires come along that, if acted upon, would thwart those desires. We tend to inhibit those desires. Sometimes the desire in question is too strong; so we gratify it and thus either suspend or altogether junk the plan. But such a situation, though all too common, represents a breakdown in the structure that mediates between desires, on the one hand, and *bona fide*actions, on the other.

Also, although these plans are to a large extent intended to gratify desires, and although that is what to a large extent they *should* be intended to do, the desires in which they have roots are not typically raw, instinctual ones. They are desires that incorporate one's values and also typically reflect previous decisions that one has made. I don't much *want* to read so and so's latest book on semantics. But my intention to become a semanticist requires that I do so. In reading so and so's book, I am acting on *a* desire that I have. But that desire is a consequence of a previous intention that I formed, and that previous intention reflects my values and my self-concept. Those values, in their turn, reflect---though they are not mere puppets of---raw instinctual desires.[***[491]***] But those values are not desires. Those values are judgments as to how to extend the hegemony of the agency within myself that polices sub-agential mental activity on my part. (Why "sub-agential" mental activity? Because that regulatory agency *is* my self. Of course, after my self comes into existence, that regulatory agency polices its own activities and thus polices agential activity.)

6.2 Donald Davison's analysis of what intentions are (continued)

Davidson identifies one's intentions with one's "strongest" desires. But the term "one's strongest desires" is ambiguous. There are two kinds of strength that a desire can have, which I'll refer to as "intrinsic strength" and "organization-dependent strength." An analogy will help us move forward.

Smith is, without any doubt, the best hand-to-hand fighter in the world. He is the best boxer, the best wrestler, and so on. So there is obviously some sense in which he is an extremely powerful human being. But none of the organizations that officially manage these sports will let him join them. So he wins no official competitions; he is given no awards; he has no titles. He has fought with the best and won without so much as breaking a sweat. And this fact is well known. But he's still not part of "the club," so to speak. So as much of a threat as he might be on the street, his fighting-prowess doesn't mean anything: it doesn't translate into money or acclaim or influence.

Jones is a reasonably good professional pugilist. He isn't remotely in the same league as Smith. (The two have fought. Jones didn't even land a punch.) But he is part of the club. So he *does* win awards for his fighting-talents, relatively modest though they are; he *does* make money; he *does* receive acclaim.

Even though Smith is a much better boxer and all around pugilist than Jones, Jones is a more powerful figure *within the boxing* world than Smith. In fact, Smith is powerless within that world, even though he is the best boxer alive.

The mind is much like the sort of organization of which Jones is a member. It's a club; and if a desire isn't a member of it, it doesn't lead to action. (It may become a member and *then* lead to action. And, as a non-member, it can have *effects.* But those effects will be symptoms, not actions, since the desire, not being a member of the club, cannot avail itself of the reigns of agency.) A desire can be extremely strong; but if it isn't part of the club, its likelihood of being converted into action is small. (I say "small," as opposed to "zero," because, if a desire is a strong enough, it is likely to become a part of one's psychological integration and, on that basis, to be converted into action---just as, given his level of talent, Smith would be likely, other things being equal, to be given a break and given membership in the boxing-federation.) And if a desire is weak, but it's deeply embedded in "the club," it's likely to be converted into action.

So in saying that an intention is one's *strongest* desire, Davidson hasn't said anything. If he's saying that it is the strongest desire that is *in one's mind*, regardless of whether it is part of one's psychological organization, then Davidson is wrong; for such organization-external desires don't have access to the "reigns of agency" and are less likely than organization-internal ones, even quite weak ones, to be expressed as actions. If Davidson is talking about organization-internal desires, he runs into the same problem. For, given two desires, the stronger one may be more out of line with the plans, self-concept, and general psychological organization, and may therefore be less likely than its weaker counterpart to lead to action. If by one's "strongest desire," Davidson is referring to that desire of one's that, *given its position in one's personality-architecture*, is most likely to lead to action, no matter how intrinsically weak it is, then his position is false or empty. For Davidson is then saying either:

> (a) That an intention is that desire that, despite very possibly having less strength than other desires, is so positioned that it will beat out those other desires;

or:

> (b) The true measure of a desire's strength is its likelihood of being expressed in action.

If (a) is the right interpretation, Davidson is simply wrong to say that one's intention is one's strongest desire. If (b) is the right interpretation, Davidson is stating the empty truism that the desire that is most likely to lead to action is the one that is most likely to lead to action.

The most fundamental problem with Davidson's analysis, however, is the one that, following Bratman[492], we identified earlier. Behaviors that directly result from desires are reactions, not actions. The immediate cause of any *bona fide*action is, not that desire itself, but the judgment-driven lifting of an inhibition on a desire.

6.3 Nietzsche's ethical egoism: An exhortation to avoid sociopathy[493]

Nietzsche infamously argued for a doctrine known as *ethical egoism* (EE). According to this doctrine, it is one's sole ethical obligation to promote one's own welfare.

One important desideratum that an ethical theory must satisfy is to say how people should treat others. Obviously EE fails to satisfy this desideratum. Since no viable ethical theory can fail to satisfy this desideratum, EE fails as an ethical theory.

But there are desiderata that an ethical theory ought to satisfy and that, unlike every other ethical theory, EE *does* satisfy. First of all, an ethical theory shouldn't require people to diminish themselves. The most pure way of being evil towards someone is to diminish them. An ethical theory that diminishes those who follow it is therefore evil. And an ethical theory that is itself evil is *ipso facto* not a coherent one. Aside from EE, all ethical theories *do* require people to diminish themselves: they require people to do what *others* want and need them to do, not what *they* want and need to do. EE doesn't require people to diminish themselves. For EE's sole injunction is: Don't diminish yourself! EE isn't evil towards those who follow it. Thus, there is a respect in which EE is coherent and in which all other ethical theories are incoherent. (That said, there are respects in which EE is incoherent and in which most, if not all, other ethical theories are coherent.)

There is another important desideratum that EE satisfies and that all other ethical theories fail to satisfy. People are self-centered. People who don't allow themselves to do what they want to do are frustrated, unhealthy people. (When I talk about "what people want to do," I mean what people want to do *in so far as* what they want to do isn't the result of cowardice or complacency or otherwise self-diminishing factors. In other words, I am talking about the wishes that people's *selves* have, and I am therefore *not* talking about the craven impulses that, although *in* people, are not *of* them---i.e., that, although in people's *minds*, are not constitutive of their *selves*.) One can obey EE faultlessly without jeopardizing one's psychological health. The same cannot be said of any other ethical theory. All other ethical theories require one to put other people's interests ahead of one's own. And, setting aside people who are *already* severely damaged, nobody can do this unremittingly. A system of morality with which it is not psychologically possible to be unremittingly compliant is *ipso facto* incoherent; for one has no obligation to do what one cannot do ("ought implies can").

"But some people may genuinely want to be kind to others," it will be

said. “EE forces people to be evil and thus to go against their own natures. So, contrary to what you say, EE *is* evil towards its own followers.” Not true. EE says: “Do what you want to do, so long as what you want to do is authentic and doesn’t involve a diminution of who you are.” Not everybody always wants to be a predator. In fact, few people are *completely* predatory, and many people are nurturers by nature. Kindness and altruism, though often given exogenous reinforcement, usually have endogenous roots. So EE does *not* demand that one hurt people whom one doesn’t want to hurt.

Bearing all of this in mind, let AET be an arbitrary ethical theory *other than* ethical egoism, and let Smith be an arbitrary person who advocates AET. Given what we’ve said, Smith is damaged if he follows AET to the letter; and supposing that Smith isn’t damaged, he’s a hypocrite, since he tells people to follow AET, even though he himself does not do so. It’s psychologically impossible for anyone who, like Smith, is not damaged, to be unremitting in his adherence to any ethical doctrine *other* than EE. Smith is therefore a hypocrite. Given that he’s a hypocrite, he has a choice:

(i) he can admit to himself that he’s a hypocrite,

or:

(ii) he can deny that he’s a hypocrite.

No one wants to believe that he’s a fraud. So Smith will probably choose (ii). (Speaking empirically, Smith *may* choose (i)—but it’s unlikely.[494])

So, given the reasonable supposition that Smith doesn’t see himself as a fraud, he’ll take path (ii). This means that he’ll rationalize and that, to that extent at least, he’ll *ipso facto* sociopathize. I suspect that, in advocating EE, Nietzsche was trying steer people away from the path of sociopathy-conducive rationalizations. Nietzsche was obviously smart enough to know that EE is not viable as a theory as to how people should treat one another. But he also said explicitly that ethical theories *other* than EE either damage their adherents or turn them into frauds or both.

Nietzsche said many things that, when taken out of context, suggest that he wanted people to be evil towards others. Presumably, we should take authors at their word. But supposing, if only for argument’s sake, that Nietzsche did want people to be evil towards one another, his wanting that

was subordinate to his wanting people to be authentic.

"Supposing that you're right," I have been asked, "why didn't Nietzsche say what you just said? Why did he not duly disclaim his exhortations that people look out for themselves? Why did not he say 'I don't *really* want you to be evil towards others; what I *really* want is for you not to be hypocrites.'?"

Nietzsche's basic message was: "people are weak and cowardly, but they tell themselves that their weakness and cowardice are virtues; and as long as they keep believing this, they cannot live authentically." Obviously one of Nietzsche's objectives in making this point was to point out an interesting fact about the human condition. But that wasn't his only or, indeed, his primary objective. His primary objective was to cut through the rationalizations by which people convinced themselves that their cowardice and weakness were virtues and by which, consequently, they kept themselves in a diminished state. Nietzsche was trying to crack the whip. Had he duly disclaimed his words, he would have failed to do that; his points would have registered with people as purely theoretical ones. Moreover, his equivocations and qualifications would have undermined his own message. For they would (rightly) have been seen as embodying the slimy evasiveness of the scholar who, not wanting to be pinned down, rationalizes his pedantry as "rigor"; and he would thus have been seen as exemplifying the very cowardice and hypocrisy that it was his intention to extirpate.

In this connection it is to be noted that Nietzsche invented the concept of "intellectualization." To intellectualize is to make a painful truth seem innocuous by treating it as being purely theoretical. Intellectualization is a form of repression. But, in intellectualizing, what one represses (or keeps repressed) is not one's knowledge of the painful truth *per se* and instead one's knowledge of how that truth bears on one's life-situation.

By keeping one's knowledge of that truth *per se* unrepressed, one is at liberty to use it in discursive contexts, and one is also spared having to rationalize away any evidence for it that one encounters. Intellectualization is thus a "thinking man's" form of repression. Like everyone else, intellectuals engage in repression. Sometimes, what they wish to repress is knowledge of a truth that, so long as they consciously know it, enhances their intellectual powers. Not wanting to forfeit the intellectual benefits of having that knowledge, they compromise: they don't repress that knowledge *per se* but they *do* repress their knowledge of its bearing on their lives. So, *qua* piece of

theoretical weaponry, that knowledge remains conscious. But, *qua* knowledge that warrants their changing themselves in fundamental ways, it is repressed.

Let us end this section by dealing with a problem that came up earlier but that we did not then have the chance to address. We saw some reason to believe that ethical theories are categorically incoherent. Supposing this to be correct, if only for argument's sake, a question arises: If we find an ethical theory to be incoherent, should we cease to endorse it? No. Any viable ethical theory is necessarily incoherent. We have a word for that incoherence: *injustice.*

Unless restrained, people will destroy the institutions that give our lives much of their meaning and value. Thus, any viable ethical theory restrains people. At the same time, any instance case of restraining someone is *ipso facto* unjust, since injustice *is* the thwarting of aspirations, even though such an injustice may be necessary to off-set some far greater injustice.

When choosing between ethical theories, we are choosing between bad and worse, not between good and better. An extensive defense of this contention is given in my (2010) book *Morality, Politics, and Law*. So I will say no more about it here.

6.4 Some famous psychopaths

I would like to end by revisiting the topic of psychopathy.

Psychopaths are always charlatans. For this reason, among others, they tend to be cult leaders. Many an "enlightened" sage is a psychopath. If someone is described as having a "spacious consciousness", odds are: he's a psychopath.

Ira Einhorn, an icon of the 1960's, was a very pure example of a psychopath. After being put on trial for murdering his girlfriend, Einhorn fled, was convicted, and sentenced to death *in absentia.* During that time, he lived in France. (At the time, France's policy was not to extradite fugitives who faced the death penalty. Hence Ira's decision to reside there.) Despite being well known in France as somebody who had been convicted of murdering his girlfriend, he was extraordinarily successful at attracting women. He managed to convince them, as he did many males, that he was the victim of a government conspiracy----notwithstanding that, by his own admission, the mummified corpse of his girlfriend had been in a trunk in his closet for two years.

Einhorn's murder of his girlfriend was a mere by-product of his psychopathy. Most psychopaths are not murderers. The essence of psychopathy isn't murderousness; it's emptiness. What made Einhorn a psychopath was the fact that he was an empty shell who hid his emptiness with a veneer of learnedness and of compassion for others. Because Einhorn was so empty, nothing meant anything to him. Therefore, killing his girlfriend meant nothing to him. In a non-psychopath, a desire to kill a loved one have to power through heavy-duty bulwarks erected by one's conscience. So there would have to be intense rage. In a non-psychopath, there is no conscience; so an act that, if committed by a non-psychopath, would indicate white-hot passion might, if committed by a psychopath, indicate nothing more than a passing whim.

While a student at the University of Pennsylvania, Einhorn's mentor was a distinguished historian by the name of Morse Peckham. Peckham was very much *not* a psychopath. Peckham was initially deeply impressed by Einhorn; he thought he had had found a protégé in Einhorn. After a couple of years, Peckham realized that Einhorn's theatrical way of speaking had blinded him to the fact that Einhorn never said anything that wasn't simply a repetition of a point that he (Peckham) had made to him.

Famous for his speeches on behalf of both environmentalism and pacifism, his advocacy of each doctrine was insincere. He used environmentalism not to enhance people, but to diminish them. In telling people that virtually everything

they did was bad for the environment, he was, in effect, telling them that *they* were bad for the environment and therefore they did not deserve to exist. His environmentalism was quite as pure an expression of homicidal rage as his murder of his girlfriend.

His advocacy of pacifism was similarly underhanded. Like many psychopaths, he advocated moral relativism ("there is no right or wrong") as well as alethic[495] relativism ("there is no truth; it's all subjective"). His advocacy of world-peace was based, *not* on the principle that certain things are objectively wrong, but on the principle that, since nothing is objectively right or wrong, no culture has the right to act aggressively towards another. What Einhorn *didn't* say is that if nothing is right or wrong there's no moral reason *not* to hurt others. (According to Marxism, the concept of right and wrong are bourgeois hogwash. This made it possible for Stalin, Mao, and Pol Pot to kill on an unprecedented scale.) Also, by advocating alethic relativism, Einhorn was wiping away thousands of years'worth of hard work, since the message he was sending was that the accomplishments of thinkers over the ages were worth nothing. This message obviously appeals to people who want to be great thinkers but either lack the requisite talent or aren't willing to do the requisite work.

The themes underlying his speeches, and often explicitly affirmed in them, were:

> (i) One's emotions are as good a guide to truth as a scientist's experiments.
> (ii) Astrology is no less worthy of study than physics.
> (iii) Commerce is inherently evil.
> (iv) Institutionalized authority is inherently evil.

These views *seem* liberal and progressive. They are not. For they are misleading ways of saying that

> (a) Hard work, personal integrity, and genuine ability mean nothing.
>
> (b) One's emotions, and thus one's prejudices, are the royal road to truth (that being why astrology, which one can practice without having ever so much as cracked a book, is no less worthy

a discipline than physics, which requires years of patient study).

(c) The appropriate attitude for a human being is the unbridled narcissism of an infant, an expression of which is a virulent hatred of anyone who frustrates his desires.

(d) There is nothing of value in the institutions that prevent us from killing one another and that require us to rise above the narcissistic pettiness to which, in the absence of external inducements, we default.

In urging acceptance of (i)-(iv), Einhorn was really urging acceptance of (a)-(d), and he was thus imploring people to understand the world in the same emotion-driven, projective manner as our cave-dwelling ancestors. Thus, Einhorn's supposed progressiveness was the purest form of *re*gressivism. Because Einhorn was representative in this respect of an entire generation of so-called liberals, people who don't want to return to a tribal mode of existence are given an incentive to *not* be liberal, even if they haven't a conservative bone in their body.

I feel that a few remarks about my personal experiences may not be out of order, so long as it's kept in mind that those remarks may well be nothing more than prejudice-driven interpretations of facts about my life that, even if interpreted dispassionately, would reveal little about the world. I personally have known many leading members of the 1960's counterculture; a very close relative of mine is one of the two or three best known astrologers in the nation. These particular individuals are among the most ill-willed and intellectually rigid people that I know. And, having attended many functions over which they presided, I have found the same to be true of their followers. For this reason, I suspect that acceptance of 1960's-style liberalism often has roots in one, or both, of the following factors:

(1) A wish to disguise, both from oneself and from others, a virulent, all-consuming ill-will,

and:

(2) A combination of a rabid desire for power, coupled with a narcissistic unwillingness to put in the work required to achieve success within institutions where they don’t make the rules and where they’ll be subjected to meaningful scrutiny.

6.4.1 Wittgenstein

Einhorn was a big Wittgenstein fan. This was no accident. Wittgenstein was a psychopath. A below average student in both high-school and college, he received a PhD from Cambridge without taking any of the standard entrance-exams, without taking a single class, and without writing so much as a single paper, apart from the TLP. He submitted the TLP as his dissertation and, on that basis, was "awarded" a doctorate in Philosophy.

Wittgenstein supposedly labored over the TLP. If this is true, his efforts were to no avail. The end result was an absolute failure. It's maximally obscure, except when what it is saying isn't trivial or obviously false. No effort is made to clarify the obscure statements or to substantiate them. Each sentence is labored and artificial and fails to flow into the next one.

Also, Wittgenstein doesn't *say* how one sentence, or one section, relates to the next. Instead, he uses an elaborate numbering system that is meant to identify the logical interrelations holding among the various statements composing the TLP. But the numbering system, though giving the TLP a much needed appearance of rigor and coherence, only adds to the obscurity. To this day, scholars have failed to produce a single interpretation of Wittgenstein's numbering system that doesn't add more incoherencies to the TLP than are already known to be there.

In Chapter 1, we consider the TLP's main contentions and saw how wrong and shallow they were. But Bertrand Russell, the world's most distinguished philosopher at the time, and still considered by many (including the present author) to be the greatest philosopher of the last 150 years or so, gave the TLP his seal of approval. (He later deeply regretted doing so.) And *voila!* Wittgenstein had a PhD in philosophy, despite Wittgenstein's previously mentioned lack of credentials and the despite the TLP's lack of merit. The TLP is discussed at length and in depth, in Chapter 1. In any case, here are a few of the many problems with the TLP that we discussed.

The *Tractatus* contains no argumentation. It consists of assertions that, on the rare occasions when they're intelligible, are very obviously false. Consider his thesis that all meaningful statements are verifiable. No sooner did philosophers read that sentence than they saw it to be subject to obvious counter-examples (e.g. "all copper conducts electricity"). To take another example: Consider this thesis that all complex statements (statements consisting of other statements) are ***truth-functions*** of the component statements.[496] In other words, given a

complex sentence, e.g. “Smith is tall and Jones is short”, whether that sentence is true is determined by whether its components (“Smith is tall” and “Jones is tall”) are true.) Again, no sooner does one read this sentence than one sees it to be subject to obvious counterexamples. (Consider the sentence: “Jones got fired because he has poor people skills.” Given only that each of “Jones got fired” and “he has poor people-skills” is true, it’s an open question whether Jones’ getting fired had anything to do with his having poor people-skills.) Given the profound differences between iconic and non-iconic forms of representation (which are discussed in Chapter 5), Wittgenstein’s “picture-theory of meaning” can only be seen as making the empty point that sentences are picture-*like* in that, like pictures, they are (sometimes) representational. With a few trivial exceptions, none of the other intelligible sentences in the *Tractatus* fares any better.

The *Tractatus* is a tiny work. If printed in the same font as the sentence you are reading, and on pages the size of the one before you, it would be around 10 pages long. But it appears much longer than it is. The reason for this is that it isn’t written in paragraphs, like other books. It consists of numbered remarks, each of which is separated from its neighbors by large spaces, as if to emphasize its importance.

The remarks are numbered by decimals, not whole numbers. The decimals are supposed to indicate relations of subordination. (So, remark #5.211 is subordinate to remark #5.21, as is remark # 5.212. I use a similar system in this book.) Thus, the numbers are supposed to indicate the logical relations holding among the corresponding statements.

But in the *Tractatus*, there is no discernible relationship between a given statement’s number and its relation to the statement, or statements, to which it is supposedly subordinate. A philosopher named Kevin Gibson wrote a paper with the title “Is the numbering system in the *Tractatus* a joke?”[497] Presumably the author thought so. I agree with him.) The numbers give the *Tractatus* an appearance of hard-hitting rigor that a reading of it fails to bear out.

At the time, Russell was the most influential and respected figure in the philosophical world. Having convinced Russell of his philosophical greatness, Wittgenstein enlisted Russell’s help in finding a publisher for the *Tractatus*. Russell was obviously as capable as anyone in the philosophical community of convincing publishers to accept a work they would otherwise reject. Evidently, the publishers judged that they had more to lose in the way of would-be publishers. (And that determination would have been correct had they been

operating in any context where book sales are driven by market-forces, as opposed to hegemony-internal politics.) Nonetheless, the first 20 publishers whom Russell urged to publish Wittgenstein's credibility by publishing Wittgenstein's work than by being black-listed by Russell. In any case, the 21st publishing house agreed to publish it, but only on the condition that Russell write a substantial preface to it, so that the book would basically be a paper by Russell, which would inevitably be good, followed by Wittgenstein's curious little effort, which, it was presumably hoped, it would overshadow.

Russell's preface is both illuminating and enjoyable to read. Wittgenstein rewarded Russell for his efforts by saying that his preface thoroughly misrepresented the *Tractatus*, but without saying how it did so and without entertaining the possibility that it might be *his* fault that nobody, including Russell, could understand what he was saying. (Russell *did* in fact misrepresent *some* of Wittgenstein's claims. But, in so doing, he attributed views to Wittgenstein that were more cogent than those advocated by Wittgenstein himself.) Wittgenstein fumed for the entirety of his career about how nobody understood what he wrote. He never blamed himself for this. It was always the reader's fault. Soon after the *Tractatus was published*, every clear statement in it, with a few unimportant exceptions, was shown to be false, thoroughly confused and, unlike other false, confused statements, not even obliquely suggestive of anything that was true or fruitful. (See Chapter 1.) Since it was written by Wittgenstein, who everybody knew to have been given Russell's seal of approval, the *Tractatus* was treated with reverence (and still is). But not a single philosopher could find anything in the *Tractatus* that, no matter how liberally interpreted, was even remotely close to the truth or that provided clear evidence that its author had any philosophical ability.

In the immediate aftermath of its publication, when the grievous problems with the *Tractatus* became public knowledge, Wittgenstein vanished for ten years, resigning his professorship and severing all contact with anyone in the philosophical community. During that time, he had no jobs of substance. He was a middle school teacher for a while, but was fired for physically abusing his students.

When he returned to Cambridge, he argued that there is no truth (or, what is the same, that truth is culture-dependent, being nothing more than a myth we create to validate our linguistic practices); that, so far as they don't consist of sensations, minds are identical with overt bodily movements; that all philosophical puzzles are immediately solved by looking at how words are

ordinarily used; and that mathematical statements never represented discoveries and were in all cases conventions or stipulations.

After returning to philosophy, Wittgenstein published nothing. He disseminated his thought through lectures. He did not permit people to attend his lectures unless he had a personal relationship with them. Distinguished philosophers from other universities were not allowed to attend his lectures. He thus ensured that the *Tractatus* would be the last work of his that, during his lifetime, was made available to anyone over whom he didn't have personal power.

Most of his lectures were transcribed. The ones that his followers judged to be the more presentable ones have been published. Once somebody who desperately wanted to see merit in Wittgenstein's work, I have read them carefully. They consist of scattered remarks. The remarks are in most cases devoid of any discernible meaning. They are seldom clarified or substantiated.

That said, certain themes clearly run through them. There is no truth; what scientists refer to as "discoveries" are just changes in the conventions governing certain terms; if one knows how a word is used, one knows everything there is no know about its meaning. Psychopaths don't *really* understand terms like "right" and "wrong," "good" and "bad." But they know how to *use* such terms. In saying that, once one knows how a word is used, one *ipso facto* knows everything there is to know about its meaning, Wittgenstein was, I suspect, trying to convince himself that, falsely, that he *did* grasp the meanings of moral terms, given that he knew how they were used (at least relative to a conception of word-usage, such as the one that Wittgenstein advocated, that takes into account public behaviors, but not their psychological underpinnings).

What all of these themes have in common is that, if they're right, everybody who has ever sought truth was on a fool's errand. The nihilistic currents that run through the various remarks composing his lectures give them a superficial unity. But ultimately those lectures are heaps of disjoint comments, which, unlike the two-liners produced by other thinkers, are uninformative and painful to read.

But those who listened to those very lectures were mesmerized by Wittgenstein's charisma. (Among those was Norman Malcolm, who, following Wittgenstein, claimed that people don't actually have dreams, our misguided belief to the contrary being a product of our mistakenly thinking that the words that people utter when discussing their (non-existent) dreams are anything more than empty noises.) It is suggestive that, setting aside self-described

Wittgensteinians, those who read his lectures have a negative reaction to them, if they have any reaction at all, whereas those who saw him talk reacted very favorably. It's also suggestive how carefully Wittgenstein vetted those who wished to attend his talks.

One point that Wittgenstein *did* make clearly is that there is no psychological component to linguistic behavior. He said that thought "consists in manipulations of symbols." He also said that linguistic symbols are inherently public. In his view, it isn't by virtue of any mental state of yours that when you say "snow is white," you *mean* what you are saying, whereas, when a parrot says it, it doesn't mean anything. Although he wrote in a sufficiently loose and obscure manner that always allowed him the option to back away from what he seemed to have said, he made remarks that either mean nothing or mean that beliefs, values, and other enduring psychological structures, are identical with the overt behaviors of which, so it is generally presumed, they are but the effects; and he made remarks that strongly suggest that he didn't believe there to be such a thing as consciousness. His ridiculous view that thought is an inherently social, as opposed to individual, phenomenon is an artifact of these clearly erroneous assumptions.

Wittgenstein's false identification of the mind with bodily behavior (plus consciousness) makes it pointless to examine oneself. It is thus an attractive view for somebody who doesn't *want* to examine himself. Wittgenstein said that "there are no mysteries." It was unclear whether he was referring to the mind or to philosophy. He seems to have been referring to both. If he were right about this, then one would have *carte blanche* to believe one's own rationalizations.

When he first met him, Russell described Wittgenstein as "the purest example of genius that I know." In so describing Wittgenstein, Russell was describing the latter's *mannerisms*, not his thought. When substantiating this characterization of Wittgenstein, he cites the fact that Wittgenstein would alternate between long silences, during which he appeared engrossed in thought, and sudden bursts of speech, which, though obscure, were accompanied by energetic, almost aggressive, gestures that made it clear that, think what one may of them, those outbursts contained true insights. Russell mentions only one comment that Wittgenstein made during his conversations with him. Russell had put forth an argument, in which it was assumed for argument's sake there wasn't a rhinoceros in the room they were in. Wittgenstein's only response to this argument: "how do you *know* that there isn't a rhinoceros in the room?" As any philosophy professor knows, this is a perfect instance of the knee-jerk

skepticism on which people with no talent for this discipline fall back when they don't know how else to engage what they're being told. [498] After knowing Wittgenstein for a month or so, Russell wrote that Wittgenstein had nothing of substance to say and that conversations with him were a "waste of time."

When Wittgenstein met Russell, Wittgenstein's father was the wealthiest man in Europe. After inheriting much of his father's wealth, Wittgenstein himself was the wealthiest man in Europe. Why mention this? Because if a billionaire shows up and asks you for a professorship, you'll have a strong incentive to give it to him (if, like Russell, you have the power to do so); and you'll look very hard for legitimate reasons to do so. And if you can find anything resembling such a reason---anything that, by any stretch, could pass for such a reason----you'll pounce on it. In any case, it seems unlikely that, if Wittgenstein had been a coal miner's son, his sudden appearance at Russell's door would have elicited such hospitality.

In his later years, Russell (1958) withdrew his support for Wittgenstein's work. Though Russell doesn't say so explicitly, he seems to have thought that Wittgenstein had played for him for a fool and that his support of him was misguided. What Russell does say explicitly is that Wittgenstein's post-Tractarian work "did not influence me at all." And what he says he says about the *Tractatus* consists mostly of arguments, all of them cogent, to the effect that specific points made in it are false. His remaining comments about the *Tractatus* express irritation (and, I suspect, a certain moral outrage) at the fact that his points are delivered without any justification, "like a Czar's ukase", and at the related fact that, because he felt his statement not to require substantiation, he didn't clarify them. Wittgenstein's writing is obscure in the extreme. When people put forth interpretations of his work, Wittgenstein *always* said that they were wrong, but never said exactly how they were wrong. There was one exception to this. Wittgenstein claimed that Frank Ramsey, a distinguished philosopher and logician who died at age 26, had correctly interpreted the *Tractatus.* I strongly believe that Wittgenstein's saying this was a strategic move. At the time, Ramsey was regarded as a luminary by both philosophers and economists. (He is still so regarded.) By saying that Frank Ramsey understood the *Tractatus*, he was saying that his work *wasn't* meaningless: it only appeared that way to people who, unlike Ramsey and himself, were too stupid to understand it. Also, given how respected Ramsey was, had Wittgenstein claimed that Ramsey didn't understand what he was saying, he would have aroused immediate suspicion, and the ensuing reappraisal of his

work would have put a swift end to his career.

Also, by claiming that Ramsey was the only person who understood his work, Wittgenstein was affiliating himself with actual geniuses, like Einstein, whose work *is* initially understood by only a few select people. Wittgenstein was also thereby affiliating himself with Ramsey, which, speaking tactically, was a very shrewd move on Wittgenstein's part. When Wittgenstein's cult of personality was still in its formative stages, it depended largely on the spurious presumption that he was a gifted mathematical logician; and by associating himself with Ramsey, Wittgenstein was able to give temporary credibility to that presumption. Wittgenstein's cult of personality was already airborne by the time it was common knowledge that the *Tractatus* contained pseudo-mathematical fatuities where it was initially believed to contain mathematical profundities.

People in Wittgenstein's inner circle didn't question the Master or ask him to clarify what he was saying. Contemporary Wittgensteinians become very uneasy around people who ask for clarification. I've noticed that every competent student that I've had was left cold by Wittgenstein's work. I've also noticed that when asked to discuss anything that was written clearly, these students of mine reacted with blank incomprehension, followed by acute embarrassment.

Wittgenstein never completed anything. The *Tractatus* is not in any significant sense a complete work. It's not really a *work* at all. It consists of loosely connected remarks, on which Wittgenstein's bogus numbering-system imposes an appearance of unity. Taken individual, those remarks, when intelligible, are false or trite. The only authors whose work he engages are those whom he personally knew; and he engages their work only in a cursory fashion. In his writings, Wittgenstein repeatedly exhorts his readers not to do philosophy. He explicitly says that his works should be the last philosophical works ever written. He discouraged would-be philosophers from going into philosophy.

Wittgenstein admitted that he rarely read the work of other philosophers. In his own writing, he never develops ideas, at least not in a clear or systematic fashion. It is left to the reader to do Wittgenstein's work for him---to interpret Wittgenstein's unintelligible remarks in such a way that they express defensible contentions and, at that point, to provide arguments for those contentions.

Wittgenstein projected an air of "unremitting integrity", as one author (not a personal acquaintance of his) put it. But it isn't clear wherein his integrity lay. His work consists of unsubstantiated statements to the effect that nothing means

anything; he was physically abusive to his students and, on one occasion, to Karl Popper, a very capable contemporary philosopher with whom he was having a conversation in which Popper identified fallacies in what Wittgenstein was saying; although he technically did hold a professorship at Cambridge, he had no clearly defined responsibilities. The philosophers who matriculated under him continued to toe his line after his death, leading in some cases to positively ludicrous assertions becoming orthodoxies (e.g., Gilbert Ryle's insistence that other people are better than oneself at knowing what one is consciously thinking).

It is often said that Wittgenstein was a "genius," but those who say this never cite any actual instances of his genius; and, although some people of great intelligence (e.g., Michael Dummett) have said it, their saying it is almost always followed by an argument, usually a cogent one, as to why Wittgenstein's views are false.[499]

Wittgenstein's professional success was based entirely on the air of mystique that he cultivated. His post-Tractarian work consisted entirely in his blocking off fruitful avenues of inquiry. He insisted, in fact, that any attempt to solve any philosophical problem was *ipso facto* misguided. In the work in which he made that claim, he said, without ever saying what he meant, that "everything is as it appears" and "there are no mysteries." He also says, without clarifying what he means, that "the purpose of philosophy is to show the fly the way out of the bottle."

One can see how one could see great depth in such statements. But one could also see how one could project depths into those statements that they don't have. In fact, there is little, if anything, in Wittgenstein's work that, when interpreted literally as opposed to projectively, has any meaning at all.

Wittgenstein's native language was German, not English. He never became fluent and always spoke with a heavy accent. This, I suspect, was much of the reason he wasn't seen for what he was. It immediately disposed people *not* to take what he was saying at face value, to assume that his obscure utterances bore hidden meanings, and, consequently, to project meanings into words that weren't there.

Wittgenstein spoke little English when he first met Russell. But, even though Russell spoke fluent German, Wittgenstein insisted on speaking English. This is because Wittgenstein knew that, had he spoken in German, Russell would have seen that his failure to say anything meaningful was a result, not of his being unable to articulate his thoughts, but of his not having any. Also, because at that

time Wittgenstein's English was so poor and because, as a result, he wasn't saying anything meaningful, Wittgenstein's words were, I more than suspect, a blank a canvass onto which Russell liberally projected his own insights, this being why, despite initially believing Wittgenstein to be an imbecile and a phony, Russell came to see great depths in Wittgenstein's broken sentences.

During his lectures, Wittgenstein would pause for unusually long periods of time. During these pauses, his face and bodily movements projected deep pensiveness along with an almost maniacally intense level of concentration. To a large extent, I suspect, these silences were Wittgenstein's way of avoiding having to say anything and, therewith, of allowing his audience-members to project into his fragmentary speech the very insights that he was unable to come up with himself. But, so far as they weren't just bogus theatrics, his long pauses and looks of consternation were, I would conjecture, expressions of the fact that he had to struggle to put his thoughts into English.

Wittgenstein's work was hollow. He refused to clarify anything that he said. He refused to concede that anyone understood anything that he said. Given its thoroughly nihilistic character and its curt dismissal of any attempt to seek philosophical truth, Wittgenstein's work would have put an end to philosophical thought, had it not been for the fact that many philosophers didn't give any credence to what he said. He failed to complete anything of substance. He earned degrees and titles without going through anything resembling the process that others had to go through to earn similar accolades. His speedy journey from below-average high-school student to holder of a Chair in Philosophy at Cambridge was based entirely on the impressions that people had of his character. (More likely, it was based on *one* person's impressions of his character. The academic world being what it is, once Bertrand Russell threw his weight behind Wittgenstein, everyone fell in line.) These are all textbook signs of psychopathy.

Wittgenstein unremittingly attempted to discourage anyone, apart from a select few, from entering into the discipline of philosophy. The few intelligible remarks that he made were glib and platitudinous. When criticisms of the *Tractatus* began to roll in, he petulantly quit his job and ceased to have any communication with anyone in the discipline of philosophy. (It is worth noting that before publishing the *Tractatus*, he described it as being the last philosophical work that need ever written. His reasons: first, there was no legitimate philosophical question to which it didn't provide a definitive answer; second, it exposed defects in the discipline of philosophy that, once seen, would

warrant a universal abandonment of it.) In his post-Tractarian work, he repeatedly said that no work in philosophy should be done after his death. He used personal relationships as a surrogate for work. Given all of this, I conclude that Wittgenstein was a psychopath.

Wittgenstein gave away his large fortune. This may seem inconsistent with the contention that Wittgenstein was a psychopath. It is not. He did this to consolidate his false veneer of legitimacy. “How could a bad person do such a good thing?” Answer: To give credibility to his false veneer of legitimacy. What drives the psychopath isn’t a desire for money. It’s a desire to construct an identity to take the place of the one he doesn’t have. Wittgenstein’s famous final words were: “Tell them I had a good life.” He did not have a good life. His life was a lie. Like his acts of charity, those words were meant to make a corrupt and bogus life seem real and, thus, worthy of emulation. (Why would somebody who had actually had a good life say that on his death bed?)

What I personally find striking about Wittgenstein is that, although he wrote in what were clearly meant to be punchy two liners, nothing he wrote goes over well if quoted. Abraham Lincoln had several brilliant two-lines; so did Yogi Berra. La Rouchefoucauld had dozens. Wittgenstein had not a single one. And not for lack of trying.

6.4.2 Foucault

Michel Foucault is philosopher who was an even more pure example of psychopathy than Wittgenstein. Wittgenstein, I believe, wasn't always a psychopath. Wittgenstein's early days, he tried to do good work. When he failed, rather than admit defeat and rather than deal with the shame he must have felt, he psychopathized. (He needed time to construct a new, bogus narrative relative to which his painful *Tractatus*-related realizations about himself meant nothing---he needed time to psychopathize---this being why he retreated into virtual solitude for ten years.)

But Foucault was a psychopath from day one. At no point in his career did he show the slightest interest in the truth. His writing is even more obscure than Wittgenstein's. And when he contracted H.I.V., he had unprotected sexual intercourse with hundreds of women, without telling them about his condition. Despite this well known fact, he is revered by professors of women's studies and other disciplines (e.g. anthropology) that are staffed primarily by women and one of whose objectives is to identify the forces responsible for the oppression of women. How can this be? Even though Foucault's work, like Wittgenstein's, makes little or no sense, certain themes clearly run through it, e.g. *Truth is a lie told by the patriarchy (and therefore doesn't exist). So-called "principles of reasoning" and "rules of logic" are machinations, devised by the running dogs of capitalism, to ensure that women and other traditionally disempowered groups continue to be dominated by moneyed Westerners.* This is a Marxist, as well as a feminist (or, at least, *apparently* feminist), version of what Wittgenstein said. (Why the parenthetical hedge? Because this message says, in effect, that women are such victims of propaganda that their powers of judgment cannot possibly be true. So Foucault's message was pseudo-feminist and anti-feminist---even though it is the very message of many self-described contemporary feminists. See Chapter 1, Section 18.0 of *Morality, Politics, and Law* for an extensive discussion of genuine vs. fraudulent forms of feminism.) This at least partly explains why many feminists are willing to overlook the fact that, in effect, Foucault tried to murder scores of women.

Another important similarity between Foucault and Wittgenstein was that each cultivated an enormous cult of personality. Had they written clearly, like Carl Hempel or Arthur Pap, they wouldn't have been seen as gurus and their work would have been evaluated in the same dispassionate way in which people evaluate the work of Hempel and Pap.

Even Russell, who to this day is still considered the greatest philosopher of the last 100 years, isn't seen as a guru; there is no cult of personality surrounding him; and those who find fault with his work never try to save it by ascribing hidden meanings to it. The reason is that respect for Russell's person is contingent on respect for his work. But with Wittgenstein and Foucault it is the opposite. It is only as far as one has respect for Wittgenstein and Foucault that one has respect for their work. Their work, taken on its own, is meaningless. After reading one page of either author, practically anyone would stop reading *unless* he were given some *ex parte* assurance that, if only he'd press on, he'd discover treasure-troves of genius. Having read most of what Wittgenstein wrote (that was published), along with the leading explanations of it, and having read a fair amount of Foucault, I can only say: There's nothing there; those epiphanous discoveries never happen. What *does* sometimes happen is that one invests so much time and energy in these authors that, rather than admit that one has been on a fool's errand for however many years and rather than admit that one's dissertation was a farce, one becomes part of the lie that perpetuates the thought-crippling myth that it's worth one's time to study what these two wrote.

Willful obscurity is not merely evidence of psychopathy; it *is* psychopathy. There are great thinkers (e.g. Kant) who write obscurely. But the obscurity is not deliberate; and, ultimately, much of what they're saying *is* clear, even though it has to be read a few times for its meaning to emerge. Had Kant written more clearly, it would be more enjoyable to read his prose, and one wouldn't have to work so hard to figure out what he was saying. But if one is willing to trudge through Kant's prose, one usually finds that it's unambiguously making a point; and that point is very often an illuminating one.

I cannot understand most of Spinoza's *Ethics.* He tried to model that work of his on Euclid's *Elements*. The result, for some reason, was a text that few people understand. But there are long passages in the *Ethics* that are as clear as a bell and that contain sparkling insights. And Spinoza's other works are clear and insightful.

Kant and Spinoza were not psychopaths. So obscurity isn't necessarily a sign of psychopathy. But *willful* obscurity *is* a sign of psychopathy, especially when it's accompanied, as it was in Wittgenstein's case and in Foucault's, by a refusal to provide clarification.

I will end this book right here. There will be no concluding remarks. Any attempt to distill the material put forth in the preceding pages into a single,

tidy conclusion would involve my imposing a false veneer of unity on that material. Which, in its turn, would involve either my lying to myself, and thus sociopathizing, or lying to you, and thus criminalizing.

Appendix: Definitions of key terms

Algorithm*:* A fixed procedure for carrying out a task The rules that we learn in grade school to multiply, add, etc., multi-digit numbers are algorithms. By *formalizing* inferences, logicians create algorithms for determining whether, given two statements, one of them follows from the other.

A problem with algorithms is that what one has to know *in order to know* whether a given algorithm is applicable to a given task is often much heftier than what one would have to know to carry out that task in an ad hoc manner. In this fact lies the doom of attempts to algorithmize—or, to use the more popular term—"mechanize" thought.

Ambiguity vs. indexicality*:* An ambiguous expression is one that is assigned meaning by two or more different rules. So "dumb" is ambiguous, since there are two rules that assign meaning to utterances of it. One of those rules is: given an utterance of "x is dumb," that utterance is true just in case x cannot speak. The other is: given an utterance of "x is dumb," that utterance is true just in case x is not intelligent.

An "indexical" is a context-sensitive expression. For an expression to be context-sensitive is for there to be some *one* semantic rule that assigns different meanings (or referents) to it, depending on the context. An example of such expression would be the pronoun "I." That expression isn't ambiguous. Some one semantic rule assigns meaning to any two tokens of it. (That rule is: given any token t of "I," t refers to the person who produced t.) But, that rule being what it is, different tokens of "I" refer to different people.

Ambiguity vs. vagueness*:* Whereas an expression is ambiguous if it's assigned meaning by more than one semantic rule, an expression is "vague" if, supposing it unambiguous, it is assigned meaning by *no* well-defined semantic rule, and if, supposing it ambiguous, one of its disambiguations is assigned meaning by no well-defined semantic rule. Someone with zero hairs is definitely "bald," and someone with a million hairs (provided that they're suitably located and have the requisite thickness) definitely is "not bald." But there are many people with an intermediate number of hairs with respect to whom neither "bald" nor "not bald" is clearly applicable. Since, therefore, the semantics of "bald" is given by a rule that is undefined for these intermediate

cases, “bald” is vague.

Like many other expressions, the word “dumb” is both vague and ambiguous, since it is assigned meaning by two distinct semantic rules, neither one of which is entirely determinate.

Commentary: There is a raging debate as to what vagueness is. Some hold that it is a property *only* of words, thoughts, and the like. Those who hold this are said to have an *epistemic* view of vagueness, and are sometimes known as “epistemicists.” (“Epistemic” means “having to do with knowledge.”) This is because they see vagueness as a property, not of the thing known, but of our knowledge of it.

Others hold that vagueness is a property of the thing known, and not, at least not merely, of our knowledge of it. Those who hold this are said to believe in “objective” vagueness, the reason being that they believe vagueness to be a property of objects, as opposed to our methods of representing them.[500] According to this view, there may be no fact as to whether Pat is a male or not or as to whether Pat is kind or not, etc. If this is right, then the law of excluded middle (any given proposition P is either 100% true or 100% false) is false.

Here is my view. There is no objective vagueness. “What’s out there is out there,” as my former colleague Chris Buford once put it. Talk of vagueness in the world is projective. We’re projecting deficiencies in our representations of the world onto the world itself. Vagueness is a property of beliefs, symbols, and other representations.

Here's why I say to this. To say of a predicate P that it's vague is to say that there is, or at least could be, some object x such that ‹Px› isn't true and ‹not-Px› isn't true. There's nothing mystifying about this sort of vagueness. Symbols have the meanings we give them. If we decide that Px is true if x is 1, 2, or 3, and false if x is 4, 5, or 6 then ‹P7› isn't true and neither is ‹not-P7.› But this isn't because there is some *proposition* that is neither true nor false. It's because, given how P has been defined, ‹P7› doesn't *encode* a proposition and neither does ‹not-P7.› For a representation or symbol to be "vague" is simply for it to be under-defined.

But what would it be for a symbol-*meaning* to be vague? But what would it be, for example, for a sentence-meaning (a proposition) to be vague?

Propositions are individuated by their entailment-relations. In other words, a proposition's identity is a function of what it entails and also of what entails it. If proposition P_1 entails P_2, but P_3 does not, or P_2 entails P_1 but not P_3, then P_1 and P_3 are *ipso facto* different propositions. Thus, the very idea of a proposition whose entailment-relations are indeterminate is a non-starter. This means that propositions always have determinate truth-conditions. After all, for proposition to have such and such truth-conditions *is* for such and such propositions to entail it. Thus, a proposition's truth-conditions are either fulfilled or they aren't. There is no other possibility. Thus, there are no indeterminate propositions, and the law of excluded middle (*every proposition is either 100% true or 100% false*) is correct. See "entailment," "propositions."

Antecedent and consequent*:* In *if P, then Q*, P is the antecedent and Q is the consequent. So in *if snow is cold, then snow is not hot*, the statement *snow is cold* is the antecedent, and *snow is not hot is the consequent.* See "conditional."

Analytic truth*:* "Squares have four sides" is analytic, since it makes no sense to suppose that squares might fail to have four sides. "Anything that is literate is animate" is analytic since it makes no sense to suppose that anything inanimate should be literate. In general, a statement is analytically true if its negation is incoherent.

See "coherent," "empirical truth," "entailment," and "meaning vs. entailment-relations."

Atomic proposition*:* A proposition that isn't molecular (e.g., *John snores*). An atomic sentence is a sentence that expresses such a proposition (e.g., "John snores"). See "molecular proposition."

A priori*:* Knowledge is a priori if one has it, not in virtue of any observations that one has made, but in virtue of one's innate cognitive structure of one's mind. See "empirical knowledge."

A posteriori*:* Knowledge is a posteriori if it isn't a priori. See "*A priori.*"

Axiom*:* A statement that, in some context, is assumed to be true and therefore isn't argued for.

Axiom (second definition): A statement-form that, in some context, is assumed to have a true interpretation. See "interpretation."

The axiom of comprehension*:* This is the principle that, given any property, there is a set containing all and only those objects that have that property. There is a set containing all and only those things that are people (i.e., that have the property of being a person). There's a set containing all and only those things that are acorns (i.e., that have the property of being an acorn). And so on.

The axiom of comprehension seems self-evident. But Bertrand Russell discovered that it has self-contradictory consequences and is therefore false. The set of people is not a member of itself, since no person is a set. But some sets do seem to be members of themselves. The set of sets would seem to be such a set.

In light of these points, consider the property of being a set that contains all and only those sets that are not members of themselves. Let S be the set that contains all and only those things that have property. Does S itself have that property? In other words, is S itself a self-member? If it is, then it isn't, since it contains only sets that aren't self-members. But it if isn't, then it is, since it contains every set that is a self-member. Thus:

(i) S has the property of being a self-member

entails its own negation, and so does

(ii) S doesn't have that property.

Given the axiom of comprehension, one of those two statements must be true. Since the axiom of comprehension therefore entails a contradiction, it is false. How exactly it must be modified is a matter of debate. See "axiom," "property," and "set."

Axiom of extensionality*:* Sets are identical if and only if they have exactly the same members.

Axiomatic system*:* The set of statements consisting of certain axioms along with the statements that follow from them. See "axiom."

Axiomatization*:* To axiomatize a discipline is to identify a small number of propositions and a small number of inference-rules such that all of the results of that discipline follow from those propositions by means of those inference rules. An axiomatization of a discipline is what results when it is successfully axiomatized. See "inference rule."

Backtracking counterfactual*:* *See "counterfactual."*

Bound variable*:* See "open-sentences."

Causal law: A law to the effect that, if one thing occurs (e.g., fire), some other thing must occur (e.g., smoke).

Causal inference*:* An inference based on knowledge of a causal law (e.g., I see smoke; I know that there is smoke only if there is fire; so I infer that there is, or was, a fire).

Causal series*:* A series of events such that, given any two of them, one is the (in)direct cause of the other; in other words, a single series of events any one of which is (in)directly caused by some one state of affairs.

Like all series, causal series don't branch. This is a tautology. A 'series' is defined to be a non-branching structure.

Of course, just a river can fork, so a causal series can branch out. In fact, the forking of a river is a special cause of the branching out of a causal series. But for a river to fork is for it to become *two* rivers (supposing that the two tributaries don't later merge). It isn't for *a* river to become two independent streams of water. Similarly, for a causal series to branch is for it to become multiple series. It is not for *a* series to comprise independent tributaries.

Class*:* Synonymous (in this work, though not in others) with "set." See "set."

Coherence and incoherence*:* An "incoherent" statement is one that undermines itself. Consider the statement:

(BC) "Bill has four cars, but he doesn't have more than one car."

BC is self-defeating. Its own meaning prevents it from being true. Therefore it is "incoherent."

It's clear that *any* statement to the effect that someone had four cars while having no more than one car would be incoherent. So the concept of such a person is one that couldn't be actualized; and any statement to the effect that it was actualized would be incoherent. Such concepts are described as "incoherent." (In this context, the term "concept" is synonymous with the term "condition," and the expression "to actualize a concept" is synonymous with the expression "to fulfill a condition.")

Concepts are conditions. To actualize a concept is to fulfill a condition.

"Ceteris paribus": Synonymous with "other things being equal" and "holding all other factors constant." *Ceteris paribus*, a person with money is more likely to be happy than a person with no money. In other words, given two *otherwise comparable* people, one of whom has money, the other of whom does not, the former is more likely to be happy than the latter. But if the former has a painful chronic illness and his dreams have been all been shattered (etc.) whereas the former is in the pink of health and he feels himself to be on the cusp of success, the latter will probably be happier than the former.

Coherently conceivable circumstance (or scenario)*:* A circumstance is

coherently conceivable just in case one is not guilty of self-contradiction in virtue of believing it to hold. So even though Kerry is not the U.S. President in 2007, the proposition *Kerry is U.S. President in 2007* is not self-undermining—it isn't like *x is knowledge but not belief* or *1 + 1 = 3.* So while all beliefs requiring the existence of circumstances that are *not* coherently conceivable are false, not all false beliefs require the existence of such circumstances. See "coherence and incoherence."

Compatibilism*:* The doctrine that there can be personal freedom in a deterministic world. "Incompatibilism" is the doctrine that compatibilism is false. See "determinism" and "incompatibilism."

Compatible*:* P and Q are compatible statements if they can both be true. Thus, P and Q are compatible if, supposing that the one is true, it doesn't follow that the other is false. "So *JM is a philosophy professor*" is compatible with "*JM is more grouchy than Steve Carey*," since they can both be true. See "entailment" and "incompatible."

Complex expressions vs. simple expressions*:* Some expressions are simple (e.g., "Smith"), while others are complex (e.g., "the man that the dog bit"). All sentences are complex—that is, they consist of multiple expressions.

At first, it appears as though there are exceptions to this. Consider the sentence "leave!" This seems simple. But it isn't; that is because there is a hidden (or "phonetically unrealized," meaning that there is no overt, acoustical, or visual representation of them in that word) occurrence of "you"—that sentence means: *you*—leave. Also, the verb leave has various semantic properties (e.g., tense, mood) that, although they are not phonetically marked, are nonetheless present, and make it a complex expression.

Genuinely simple expressions are known as "morphemes."

Compositionality*:* According to the *principle of compositionality*, the meaning of a complex expression is a function of the meanings of its constituents; and the referent of a complex expression is a function of the referents of its constituents.

The meaning of:

(TY) “the author of *War and Peace* was friendly to the author of *Crime and Punishment*”

is different from that of

(DO) “the author of *War and Peace* was mean to the author of *Crime and Punishment.*”

Why do TY and DO have different meanings? Because “friendly” and “mean” don’t mean the same thing. Replacing “friendly” with “mean,” or *vice versa*, turns a sentence having one meaning into a sentence having a very different meaning. This shows that what TY means *depends* on what “friendly” means. Semanticists express this by saying that TY’s meaning is “a function of ” what “friendly” means. What DO means is a clearly function of what “mean” means. In general, what expressions mean is a function of what the expressions composing them mean. See “special compositionality,” “function,” and “functor.”

Compound sentence*:* Same as a molecular sentence. Thus, a compound sentence is one that consists of other sentences (e.g., “snow is white and grass is green”). See “molecular sentence.”

Conceptual role semantics (CRS)*:* the view that meaning is use, i.e., for two expressions to mean the same thing is for them to be used in the same way and, therefore, that for a given expression to have this, as opposed to that, meaning is for it to be used in this, as opposed to that, way.

In his (1963) paper “Some reflections on language games,” Wilfrid Sellars clearly puts forth a version of CRS. But it’s quite definitely present in Wittgenstein’s book *The Philosophical Investigations*, which was completed in the late 40’s. In his (1994) book *Making it explicit*, Robert Brandom advocates CRS. Nowhere in that 900 page book, or in any writings of advocates of CRS, is the existence of any of the following reasonable positions so much as acknowledged, let alone addressed:

(i) A knowledge of a word’s meaning is at least part of what guides a

person to use it in the way she does. When the waiter asks you what you would like to eat, you say “I’d like a bowl of clam chowder,” and not “I’d like a bowl of rotten entrails.” Why do you say “clam chowder” instead of “rotten entrails”? Because you know that, given what it is that you want, “rotten entrails” has the wrong meaning and “clam chowder” has the right meaning. Of course, how an expression is used is likely to *change* its meaning. If enough people use the word “restive” to mean *tired*, it may well come to have that meaning. But until that time comes, those people are *mis*using that expression; and their usage of it, being discrepant from its actual meaning, can’t possibly be constitutive of it.

(ii) Mere noises aren’t linguistic expressions. It is only noises coupled with meanings that are such expressions. In a world where there are no semantic conventions, a noise just like the noise you make when you say “snow is white” would mean nothing. If, in such a world, the gases leaving a volcano happened to make that noise, they would mean nothing. (In fact, they’d mean nothing in *our* world, except at most in some derivative sense.) Since, therefore, noises must be coupled with meanings in order to qualify as linguistic expressions, it makes no sense to say that what an expression means is a function of how it is used. After all, there is no expression to be used *until* the noises (and ink-marks and so forth) have been coupled with meanings. Therefore, CRS is false.

Of course, there are different views as to what meanings are, and there are different views as to what it is for noises to be coupled with them. But relative to any such view, CRS is false.

The contention that is closest to CRS that isn’t obviously false is the trivial contention that what noises (and ink-marks and the like) mean is a (partial) causal consequence of what people do with them. (This suggests that CRS is based on a failure to distinguish the concept of a logical consequence from that of a causal consequence.)

(iii) Expressions that haven’t ever been used have determinate meanings. There are infinitely many sentences of English that haven’t been used but that have determinate meanings. If this weren’t the case, then nobody could ever utter a new sentence of English, at least not without making some ad hoc extension to the English language. But previously unuttered sentences are uttered all the time; and no sooner are they uttered than they have determinate

meanings.

According to CRS, what an expression means is determined by how it is used. So if CRS is right, an expression that hasn't been used *ipso facto* has no meaning.

To parry this, advocates of CRS would have to say the following. "Where some expressions are concerned, meaning is fixed by use. But where other expressions are concerned, use *doesn't* fix meaning, at least not single-handedly. Some *other* factor is at work."

In making this concession, the advocate of CRS is admitting that CRS is false. Setting this aside, let us ask: what is this other factor?

The answer: syntax. The meaning of "Bob kicked the can" is a function of (i) what each of its components means (in other words, it's a function of what is meant by each of "Bob," "kick," and so on); and it's also a function of (ii) the syntactic rules of English (in other words, it's a function of the rules that assign meanings to complex expressions on the basis of the meanings of their constituents).

Thus, the meaning of a complex expression is *not* determined by its use, and is determined by existing semantic and syntactic conventions. So CRS is false, if taken as an analysis of complex expressions. Earlier, we saw why it's false if taken as an analysis of simple expressions.

(iv) If the meaning of a sentence S were determined by its "conceptual role"—that is, by what people inferred from S, when they thought it correct—any inference that people made from S would *ipso facto* be constitutive of S's meaning. If enough people thought that "Stalin the dictator of the Soviet Union from 1923–1953" entailed "somebody who was over 6 ft tall was the dictator of the Soviet Union from 1923–1953," the former would entail the latter. But even if the latter were correct (which it isn't since Stalin was barely over 5 ft tall), it wouldn't *follow* from the first sentence.

In their (2001) book *The Compositionality Papers*, Jerry Fodor and Ernie Lepore put forth several blistering criticisms of CRS. In that work, any gaps in the criticisms of CRS just put forth are very much filled in.

Conditional*:* A conditional statement is one that has the form *if P, then Q.* Conditional statements play a vital role in logic, since logic is the discipline that attempts to determine what follows from what, and any statement to the

effect that one statement follows from another is expressed by a conditional.

Confirm*:* For P to confirm Q is for P to provide support for Q. (P: 'Smith's fingerprints are found in the murder-weapon used to kill Jones.' Q: "Smith killed Jones." P supports Q.) Confirmation is defeasible, meaning that, even if P confirms Q, some other statement might provide more support for not-Q than P provides for Q. (R: "There is a video-tape of Brown killing Smith, then pressing the hand of Smith, who is at the crime-scene but unconscious, around the murder-weapon, and then leaving it at the scene.") ***Entailment*** is not defeasible. A consequence is that P cannot, if true, entail Q without Q's being true. See "confirmation."

Confirmation*:* P confirms Q if, supposing that P is true, it's possible that Q is false but it's more likely than not that Q is true. "Smith is a Harvard professor" confirms "Smith is not a complete moron," but it doesn't entail it. See "confirm" and "entailment."

Conjunct*:* In *P and Q*, each of P and Q is a "conjunct." So *snow is white* is one of the conjuncts of *snow is white and grass is green.* See "conjunction."

Conjunction*:* A conjunction is any sentence composed of two sentences that are joined by an "and" or a "but" or an "although"—or any other connective whose purpose is to indicate that *both* of the sentences thereby joined are true.

Connective*:* A "connective" expression is one that, given one or more sentences, enables a new sentence to be formed. So, "if " is a connective since, given the sentences "grass is green" and "snow is white," it yields the sentence "if grass is green, then snow is white." All conjunctions ("and," "or," "but") are connectives. So is "not." This is because "snow is not white" says the same thing as "it is not the case that snow is white," which is what results when the expression "it is not the case" is coupled with "snow is white."

"It is possible" (or "possibly") is a connective, since it yields "it is possible that John is home" when given the sentence "John is home."

"Sally believes" is a connective since it yields "Sally believes that John is

home" when given the sentence "John is home."

"And," "if," and "or" are *two*-place connectives, meaning that they form sentences out of pairs of sentences. "Not," "it is possible that," and "John believes that" are *one-place* connectives, since they form sentences out of single sentences.

Commentary*:* A connective can be thought of as a sentence-level adjective. That is, it can be thought of as an expression C such that, given a sentence S, C yields a new sentence C(S) (read: "C of S") such that C(S) attributes some characteristic to the proposition (or open proposition) expressed by S. For example, "Fred knows that" is a connective. And if you say "Fred knows that snow is white," you are attributing a property to the proposition that snow is white; you are saying of that proposition that it has the property that Fred knows it to be true. "Because" is a connective, and if you say "snow is white because grass is green," you are seeing of the propositions that snow is white and grass is green that the former is a causal consequence of the latter. One last example: "for all x" is a connective, and if you say "for all x, x is a mammal if x is a whale," you are attributing a characteristic (that of being correct) to each of *infinitely many* propositions. (That is why, even though "for all x" and other quantifiers don't operate on sentences, they nonetheless qualify as connectives.

Connectives attribute properties to propositions. This is what quantifiers do. In fact, quantifiers do it on an even larger scale than other connectives. Consider the quantifier "for all x." When coupled with the open sentence "x is tall," this yields a sentence ("for all x, x is tall") that says of an infinitely large class of propositions (viz. those having the form x is *tall*) that each of its members is true. (By contrast, "Fred believes that," when coupled with "Jim is tall," yields a sentence ("Fred believes that Jim is tall") that attributes a property (that of being believed by Fred) to only *one* proposition.) That is why, even though quantifiers operate on open sentences, whereas other connectives operate on *actual* sentences, quantifiers are generally treated as connectives.

A sentence can have more than one connective. An example of such a sentence would be:

(*) "either Jim is in Idaho and Larry is in Idaho with Jim, or Larry is in Delaware and Fred is with Larry in Delaware."

The main connective of (*) is “or.” Why is it the *main* connective? (*) attributes a property to the set containing the propositions *either Jim is in Idaho and Larry is in Idaho with Jim* and *Larry is in Delaware and Fred is with Larry in Delaware*. The property that it attributes to that set is the property of containing at least one true member. The other connectives occurring in (*) *don’t* attribute properties to that pair of propositions; they attribute properties to propositions composing the propositions composing (*). That is why none of those connectives is the main connective. I leave it to the reader as an exercise to produce an explicit definition of the term “main connective” on the basis of these remarks.

Consequentialism*:* The doctrine that it is entirely in virtue of what its effects are that an act is morally right or wrong. See “deontology” and “utilitarianism.”

Consistent/inconsistent*:* If two statements are compatible, they are “consistent” with each other; otherwise they are inconsistent with each other. “Consistent” and “compatible” mean the same thing. See “compatible.”

Consistent/inconsistent (in connection with axiomatic systems)*:* When an axiomatic system is described as “consistent” what is meant is that, given any two statements composing it, they are consistent with each other. An axiomatic system is “inconsistent” if it doesn’t satisfy this condition. See “axiom” and “axiomatic system.”

Content*:* The content of a statement or a belief is what must hold for it to be correct.

Content-externalism*:* Let X and Y be two creatures that are qualitatively identical *except* for the fact that the causes of X’s current condition aren’t qualitatively identical with those of Y’s current condition. According to content-externalism, the content of X’s thoughts and perceptions may, in virtue of that fact, be different from the contents of Y’s corresponding thoughts and perceptions. See “semantic externalism.”

Contingent*:* A proposition is contingent if there are coherently

conceivable circumstances where it is a true and also conceivable circumstance where it is false (e.g., *Mary owns a BMW*).

Counterfactual*:* a statement of the form: *if P were the case, then Q would also be the case* (e.g., *if Kennedy hadn't been assassinated, he would have been reelected in 1964*).

A backtracking counterfactual is one of the form: "if P were the case at time t, then Q would have been the case at time t*, "where t* is *earlier* than t." For example, "if Jim now had alcohol on his breath, then it would have been the case that he had been drinking earlier today."

Counterfactual analysis of causality*:* The position that x caused y to occur just in case, if x hadn't happened, y wouldn't have happened. According to this view, for the button's being pushed to cause the elevator to come is for it to be the case that the elevator wouldn't have come if the button hadn't been pushed. In Chapter 17, it is shown why this view is false.

Contextual definition*:* See "denotative vs. ostensive vs. descriptive vs. contextual definition."

Defeasible*:* A tendency or principle is "defeasible" if it can be overridden. A *ceteris paribus* statement is one to the effect that, for some *defeasible force F* and some initiate state of affairs S, there is some possible state of affairs S* such that (i) if F is the only force operating on S, then S* will be the output and (ii) if F is not the only force operating on S, then, for some state of affairs S#, S# will be the output and the degree to which S# approximates to S* is directly proportional to the degree to which F is dominant among the operative factors.

Definite description*:* Expressions of the form ‹the phi›, where phi is in the singular, are known as "definite descriptions." So "the inventor of Velcro," is a definite description, as is "the inventor of bifocals." See "quantifier."

Define/definition*:* To define an expression is to say what it means. Expressions are defined, not objects. "Bill Clinton" is defined, not Bill Clinton.

Definition by abstraction*:* In some cases, it can be said what it is for two things, x and y, to share some property P without using any expression that refers to P. In such cases, it is possible to translate statements in which expressions referring to that property occur into statements in which such expressions don't occur; and when this is done, the property in question is thereby defined in abstraction.

For example, two lines have the same direction (i.e., share the property of pointing in a given direction) if they don't intersect. Thus, the direction D of a given line L may be defined as the class of all lines L* that don't intersect with L, and "y has direction D" may be seen as saying the same thing as "y doesn't intersect with L."

Demonstrative*:* Some indexicals must be accompanied by an act of demonstration—an act of pointing, or some equivalent—if they are to succeed in picking out their target. If, while in the presence of 10 people, I say to you "that person is a bore," what I've said is ambiguous. But if, under the same circumstances, I utter the same sentence while pointing to Charlie, then my utterance of "that person" picks out Charlie, and my utterance is not ambiguous. An utterance of "that person" refers to the person who is salient in the context of utterance. There is thus a *rule* that assigns referents to tokens of "that person," that rule being *if "that person" is uttered in a context in which x is a uniquely salient person, then that utterance picks out x.* What that refers to is therefore systematically context-sensitive, that being why it's an indexical. And sometimes an act of pointing is needed to make somebody be uniquely contextually salient, that being why it's a demonstrative, as opposed to a mere indexical. Other examples of demonstratives are "this pain," "these little creatures," and "that aardvark." See "indexical."

Denotative vs. ostensive vs. descriptive vs. contextual definition*:* For an expression to be defined *ostensively* is for it to be defined by means of an act of pointing. If, having been asked who "Ludwig" refers to, I say "*that guy*," while pointing to Ludwig, I've defined "Ludwig" ostensively.

Ostensive definitions are always cases of *denotative* definitions, but not *vice versa.* For an expression to be defined denotatively is for it to be defined by identifying some entity that it picks out. This is how proper names are defined. "Socrates" is defined denotatively—you learn what "Socrates"

means by being told who it refers to. But since he's no longer around, it isn't defined ostensively.

Not all expressions can be defined denotatively, and such expressions must be defined *contextually*. "Something" doesn't pick anything out. Therefore, it can't be defined denotatively. Its meaning is given by the rule that, for any phi, ‹something has phi› is true exactly if phi is instantiated. Other examples: The meaning of "every person" is given by the rule that, for any phi, ‹every person has phi› means that the property of being a human non-phi isn't instantiated. The meaning of "not" is given by the rule that, for any sentence S, not-S is true just in case S is false. The meaning of "no aardvark" is given by the rule that, for any phi, ‹no aardvark has phi› is true exactly if the property of not being both an aardvark and a phi is universally instantiated. See "define/definition."

Commentary—A principle concerning contextual definition: Whenever an expression is defined contextually, as opposed to denotatively, whole sentences containing it must be reparsed if their meanings are to be made clear. We just saw this with "something." Its meaning is given by the rule that, for any property psi:

(SP) ‹something has psi›

means

(PS) the property of being a psi is instantiated.

PS doesn't have the same form as SP. PS is a wholesale reparsing of SP; and the reason is obviously that "someone" is defined contextually. The same is true of all expressions that are defined contextually.

For any psi, ‹something has psi› attributes a property to another property, and not to an individual. (In analytic philosophy, the term "individual" refers, not only to people, but to *anything* that isn't itself a property—to rocks, trees, galaxies, etc.) "Someone is tall" attributes the property of being instated to the property of tallness.

Additional commentary: All denotative definitions are contextual definitions, but not *vice versa.* To say that, for some object x, "Smith" refers

to x is to say that, for any phi, ‹Smith has phi› is to the effect that x has phi. So in defining "Smith" by identifying its referent, one *is* giving its meaning by saying what is meant by whole sentences of the form ‹...Smith...› This principle is defended in Chapter 7.

Denote*:* As we'll use them, the terms "denote" and "refer to" are synonymous, the same therefore being true of "denotation" and "referent" are synonymous. See "refer."

Deontology*:* The doctrine that it is in virtue, not of what its effects are, but of whether it embodies due

regard for one's ethical duties and for other people's rights that an act is morally right or wrong. See "consequentialism."

Determinism*:* The doctrine that *nothing* is uncaused. Everything that happens/has happened/will happen *has to happen* and, moreover, *has to happen in the exact way in which it happens.*

Alternate definition of determinism: how the world is at any point fixes how the world will be, in every respect, at all later times. See "indeterminism."

Descriptive proposition*:* A proposition is descriptive if it isn't normative. In other words, a proposition is descriptive if it, supposing it is correct, says how things are, not how they should be. *Killing is frequent, especially during war* is not a normative statement, since it merely says how things are, not how they should be. See "proposition" and "normative proposition."

Disjunct*:* In *P or Q*, each of P and Q is a "disjunct." See "disjunction."

Disjunction*:* A disjunction is any sentence composed of two sentences that are joined by an "or." For example, *either it's raining or somebody turned the sprinkler on.*

Duhem-Quine thesis*:* The thesis, which was advocated by W.V.O. Quine (1908–2001)[501], that no theory does either a better job or a worse job than any other theory of modeling any given body of data. In other words, given any body of observations B and given any two theories T and T*, T does no

better and no worse a job than T* of modeling B.

Quine's argument was that, given any datum D that appears to confirm T and to disconfirm T*, one can make D disconfirm T and disconfirm T* by tinkering with one's background assumptions.

Quine's thesis is either empty or it's false. Scientific theories aren't analytically entailed by the observations that support them. Induction isn't deduction. This is an obvious triviality. So, assuming that what Quine is saying isn't trivial, it must be to the effect that no theory does a better job or a worse job than any other theory of accounting for any given body of data. To account for a body of data is to show that the occurrence of each datum composing it is non-anomalous. Thus, Quine's thesis, supposing it non-trivial, is to the effect that no theory does a better job than any other theory of removing causal anomalies from any given body of data.

Thus interpreted, Quine's thesis is clearly false. The acne of each of the millions of acne-sufferers who takes medication X gets ten times worse within ten minutes of the time at which said acne-sufferer takes X. Theory T entails that anyone who has acne who takes X will immediately be cured of his acne forever. Rival theory T* entails that theory T is false. According to Quine, by tinkering around with background assumptions, T models the observational data as well as T*. But that's clearly false.

Quine says that, in situations like this, we should "plead hallucination." In other words, we should assume that the observations in question, most of which are had by people who do not otherwise have any history of hallucinations, are hallucinations. If we make this assumption, we're stuck with a huge number of otherwise non-existent anomalies. We're stuck having to explain, or to regard it as incapable of explanation, why it is that people who never had hallucinations before all of a sudden started hallucinating; and we are also stuck having to explain/regard as inexplicable why it is that people who did have hallucinations before all of a sudden started having acne-related hallucinations that, in terms of their content and in terms of their etiology, are completely unlike the hallucinations they had previously.

Empirical knowledge, analytic knowledge, synthetic, a priori, a posteriori*:* Empirical knowledge is knowledge that is based at least partly on the testimony of the senses. Not all knowledge is empirical; some is *analytic.*

Analytic knowledge is knowledge that is arrived at strictly through the analysis of concepts.

Synthetic knowledge is non-analytic knowledge.

A priori knowledge is knowledge that is constitutive of our cognitive machinery; it's knowledge that is hardwired into us and that we therefore don't acquire.

A posteriori knowledge is knowledge that is not a priori.

Empirical truth*:* An empirical truth is one that can be known only through sensory observation. It is only on the basis of sense-perception that one can know that Barack Obama is currently the U.S. President. That is why "Barack Obama is currently the U.S. President" expresses an empirical truth. See "analytic truth."

Empiricism*:* The doctrine that all knowledge is derived from sense-perception. See "rationalism" and "sense-perception."

Commentary: Many hold that it's "scientific" to accept empiricism and "unscientific" to reject it. If, by "scientific," one means "indicative of an inability to grasp principles of any sophistication," this is correct. Otherwise, it's false. If all knowledge were observation-based, knowledge of that very fact couldn't be observation-based. In order for observation to tell you anything, it must be presupposed that observation is a source of knowledge. Therefore, observation can tell you that observation is the only source of knowledge *only* if it's known independently of observation that observation is the only source of knowledge. So, were it known that all knowledge was observation-based, it wouldn't be a truth that all knowledge was observation-based. Therefore, empiricism, if true, is either false or it cannot be known to be true. Therefore, empiricism is incoherent, and therefore false, or false. Therefore, it's false.

Encode*:* Anything that has a meaning is said to encode it. The content of a perception (i.e., what it tells you) is what it is "coded into it" (i.e., is what is encoded in it).

Entailment/entails*:* P entails Q just in case it's impossible for Q to be false if P is true. Thus *Smith has five cars* entails *Smith has more than one car*, since it isn't possible for the second to be false if the first is true.

P entails Q if P is incompatible with not-Q. See "incompatible."

Entailment (alternate definition)*:* P entails Q just in case there is no coherently conceivable circumstance where P is true and Q is false. So *x is a triangle* entails *x has three sides* because there is no coherently conceivable circumstance where the first is true and the second is false. See "coherently conceivable circumstance."

Entailment (Second alternative definition)*:* P entails the *negation* of Q just in case P and Q are incompatible. See "compatible."

Epistemic operator*:* An *epistemic* operator is one that, when joined with a sentence, yields a sentence that describes somebody's beliefs or thoughts or feelings about the proposition expressed by the original sentence. Thus, "Jerry believes" is an epistemic operator, since, when placed in front of "that 1 + 1 = 2," yields:

(iii) "Jerry believes *that 1 + 1 = 2*,"

which describes Jerry's attitude towards the proposition meant by "1 + 1 = 2." And "Laura fears that" is an epistemic operator since, when placed in front of "*that law school will be intolerably boring*" yields

(iv) "Laura fears *that law school will be intolerably boring*,"

which describes Laura's attitude towards the proposition meant by "that law school will be intolerably boring."

Any sentence containing an operator also contains an expression denoting a proposition; and, in that sentence, that operator is attributing some property to that proposition. (iii) contains the operator "it is possible that." It also contains "that 1 + 1 = 2," which denotes a proposition. And (iii) ascribes a property (that of being possible) to that proposition. See "modal operator" and "operator."

Equivalence*:* P and Q are equivalent if each entails the other. This means that there is no coherently conceivable situation where the one holds but the other does not. So *1 + 1 = 2* and *triangles have three sides* are equivalent. Note: two statements that are equivalent may have little or nothing to do with

each other in terms of content. We'll revisit this point later. See "equivalent."

Equivalent*:* Two statements are equivalent if they entail each other. P and Q are equivalent iff P ↔ Q

Commentary: Equivalent statements cannot differ in *truth-value;* in other words, they are both true or both false (one can't be true while the other is false). But non-equivalent statements can have the same truth-value. ("JMK is a U.S. citizen" is true, and so is "JMK is a philosopher professor." But they're not equivalent.).See "entails."

"Exactly if"*:* A synonym of "iff" (which is short for "if and only if") and "exactly if" and "just in case." See "iff."

Existential generalization (first meaning)*:* It's obvious that some sentences have the form "there is at least one thing that has property phi." Here are some examples:

(i) "There is at least one thing that is a human that plays golf,"

(ii) "At least one number greater than two is an even prime."

(iii) At least one person wrote a book called *War and Peace.*

(i)–(iii) are known as *existence-claims*, since each says that there exists a thing of a certain kind. Another term for "existence-claim" is *existential-generalization.* Existence-claims don't always begin with "there is." Often they begin with "something" or "someone." Thus:

(iii*) "someone wrote a book called *War and Peace*."

is an existence-claim.

The differences between (iii) and (iii*) aren't relevant to what we're doing. As far as we're concerned, they're the same statement.

Existential generalization (second meaning)*:* An inference of the form "*O has phi; therefore, something or other has phi,*"—for example, "Bob is

over 7-feet tall. Therefore, something is over 7-feet tall."

Export/exportation*:* To "export" an expression occurring in some sentence is to move it from the inside of that sentence to the outside. Consider the sentence "snow is not white." That is equivalent with "it is not the case that snow is white." In the latter sentence, the negation-sign has been pushed to the outside of the sentence. For this reason, the second sentence, unlike the first, makes it clear that the ***scope*** of the negation is (in each sentence) an *entire* sentence (viz. "snow is white"). Making the meanings of sentences clear almost often involves expression-exportation.

Expression-types vs. expression-tokens*:* A given word can occur on several occasions. There have been countless utterances of the word "snow." An utterance (or inscription) of that word is called a *token* of that word. The word itself is called a "word-type." So word-tokens are bits of noise, ink, etc. Word-types are things of which these noises (etc.) are instances.

"Exactly if": See "iff."

Epistemically possible*:* A statement S is epistemically possible iff, given *only* what you know, it isn't ruled out. So if, given only the data at your disposal, it cannot be ruled out that Smith is not in Greenland, then "Smith is not in Greenland" is an epistemic possibility. It's obvious that epistemic possibility is a relative notion. At a given time, a given statement may be epistemically possible to some and not epistemically possible to others; and a given statement may be epistemically possible to a given person at one time but not at some other time. (If you go to Greenland, and see that Smith is there, then it's no longer epistemically possible—for you—that Smith is not in Greenland.)

Equivalent*:* Two statements are equivalent if each entails the other. See "entails."

Form of an expression*:* The sentences:

(i) "Sally hates Bob"

and

(ii) “Bob adores Amanda”

have the form

(iii) x bears relation R to y.

This is because (i) is what results when the variables in (iii) are replaced with constants, and so is (ii). In general, for two sentences S_1 and S_2 to have the same form is for there to be some open-sentence S_3 such that each of S_1 and S_2 is what results when the free variables in S_3 are replaced with constants.

Formal entailment*:* See “formal truth.”

Formal truth*:* Formal truth is a property of sentences.[502] A sentence is formally true if every sentence having the same form is true. So “if Bill is tall, then it is not the case that it is not the case that Bill is tall” is formally true, since every sentence of the form “if S, then not S” is true. That sentence is also an example of a **formal entailment.** S_1 formally entails S_2 if the sentence “if S_1, then S_2” is formally true.

Formalization*:* To formalize a discipline is to identify a set of open-sentences such that every proposition belonging to that discipline is either an instance of one of those open-sentences or is a formal consequence of some interpretation of those open-sentences.

Free variable*:* See “open-sentences.”

Function*:* A rule that assigns no more than one object to each object in some specified class of objects. Thus, the rule that assigns the number two to the number one, and the number four to the number two, etc., is a function. (In other words, F(x) = 2x is a function.) But, importantly, not all functions involve numbers—a fact that Frege turned to good account.

Functionalism*:* The doctrine that for x to be a mental state of a given kind (e.g., for it to be a belief that snow is white) is for it to have certain effects and certain causes. (More formally: given any mental category M, there are certain causes and certain effects such that, given any entity x, x falls into M iff x has those causes and those effects.) Thus, anything that has the causes and effects of (say) a belief that snow is white is itself, for that very reason, a belief that snow is white. See "materialism."

Hedonism (first meaning)*:* The psychological doctrine that nobody can do anything other than pursue his or own pleasure. One problem with this doctrine is that not all forms of well-being are identical with the experiencing of pleasure. In fact, not all forms of *enjoyment* are identical with pleasure. I enjoy playing tennis. I experience pleasure when I have a few beers. My agency is implicated in the enjoyment I derive from tennis. The joy I derive from tennis is joy that I *earn*. By contrast, the joy I derive from alcohol is not joy that I earn. My agency isn't involved. (My agency is involved in lifting the bottle to my mouth. But whereas that act is only a means to the actual source of pleasure—namely, the biochemical reactions that take place in my liver—the act of hitting a good backhand is *itself* a source of joy; it is constitutively, as opposed to merely instrumentally, involved in the joy I experience when I play tennis.) The word "pleasure" denotes enjoyment of this (non-agential) kind. The contempt that puritans have for pleasure is rooted in the fact that, whereas one's agency and, therefore, one's self are implicated in the joy that (for example) playing tennis brings, they are not implicated in the joy brought by narcotics-use and (with some qualifications) by sexual activity. So even though it probably isn't psychologically healthy to have such a severe attitude towards such pleasures, the puritan's view of them embodies the important insight that, so far as one seeks pleasure, one seeks an effacement of self. Thus, a certain respect for human beings (*qua* agents) is inherent in the puritan's view, and a respect for people (*qua* agents) is *absent* from the hedonist's view. Where the puritan goes wrong is in his failure to see that a psychobiological prerequisite for mental health, and thus for retention of one's agency, is a certain amount of passive gratification.

Hedonism (second meaning)*:* The ethical doctrine that nobody ought to do anything other than pursue his or own pleasure.

Higher-order property*:* A property of a property. The property of being a bird has many properties. It has the property of being such that there are instances of it. In other words, it has the property of being *instantiated.*

The property of being a golden mountain does *not* have that property, since there are no instances of it. It thus has the property of being *uninstantiated.*

Instantiatedness and un-instantiatedness are the two most important higher-order properties.

The property of being a person has the property of being such that each of its instances also has the property of being a mammal.

Many sentences (propositions) that appear to be about individuals are in fact about properties. Frege discovered this, and it is undoubtedly the most important insight in the history of semantics. The expressions "exist" and "doesn't exist" refer to higher-order properties. So "even numbers exist" means: *the property of being an even number is instantiated.* Thus, when appropriately re-parsed, "exist" proves to mean *is instantiated* and "even number," in that context, proves to mean *the property of being an even number*.

And "even prime numbers greater than two don't exist" means: the property of being an even number greater than two is un-instantiated. Thus, when that sentence is duly re-parsed, "don't exist" means *is un-instantiated*, and (in that context) "even prime numbers greater than two" means: *the property of being an even number greater than two.*

"Iff," "just in case," "exactly if," "↔'*:* "Iff " abbreviates "if and only if." "P iff Q" means that P and Q are equivalent. "just in case" and "exactly if" are synonyms of "iff." So "P just in case Q" means the same thing as "P iff Q," and each means the same thing as "P exactly if Q"; and each of those means the same thing as "P → Q."

Incompatibilism*:* The doctrine that there cannot be personal freedom in a deterministic world. "Compatibilism" is the doctrine that incompatibilism is false. See "determinism" and "compatibilism."

Incompatible*:* P and Q are incompatible if they can't both be true. Thus, P and Q are incompatible if, supposing that the one is true, the other is false. So *JM is a philosophy professor* is incompatible with *JM is not an instructor*

of any kind, since they cannot both be true. See “compatible.”

Identity of Indiscernibles*:* If x has every property that y has, and vice versa, then x is numerically identical with y.

Indeterminism*:* The doctrine that some things are uncaused. See “determinism.”

Indexical*:* An indexical is an expression whose referent depends in a systematic manner on the context of utterance. (More precisely, an indexical is an expression-type E such that the semantic rule for E is to the effect that what a given token of E refers to is a function of some fact about the context of utterance.) If, on April 25, 2009, I say “today it’s sunny,” my utterance of “today” picks out April 25, 2009. If, on Feb. 4, 2024, I say “today it’s not sunny,” my utterance of “today” picks out Feb. 4, 2024. (That is why those utterances don’t contradict each other and, therefore, can both be true.) In general, an utterance of “today” refers to the day on which that utterance occurs. Thus, what such an utterance refers to depends in a principled (rule-governed) and therefore systematic way on some fact about the context. Thus, “today” is an *indexical.* Other examples of indexicals are: “yesterday,” “now,” and “here.” See “ambiguity.”

Indicative conditional*:* A conditional that isn’t a counterfactual conditional. In other words, the antecedent of an indicative conditional isn’t assumed to be false (though it isn’t assumed to be true). “If Smith is at the airport, then he’s probably at the baggage carousel” is an indicative conditional, since it is left open whether or not Smith is at the airport. “If Smith *had* been at the airport, he *would* have been at the baggage carousel” is a counterfactual conditional, since it is being presumed that Smith was *not* at the airport. See “conditional” and “counterfactual.”

Indiscernibility of Identicals*:* Same as Leibniz’s Law (if x has every property that y has, and *vice versa*, x and y are numerically identical). See “numerical identity.”

Individual*:* In analytic philosophy, the word “individual” has two meanings. First meaning: An individual is anything *about which* statements

are being made. Rocks, trees, and lizards are individuals in a given context, provided that, in that context, statements are being made about them. If, in that context, no statements are being made *about* the property of being green, then, in that context, that property doesn't qualify as an "individual." But in other contexts it might so qualify. Second meaning: Individuals are discrete, spatiotemporal entities—they are "things," in the narrowest sense of the word. For reasons that are given in Chapter 16, this means that an individual is any ***causal series.*** See "causal series."

Informal analytic truth*:* A statement is analytically true if it's non-empirically true (i.e., if its negation is prohibited from being true by the structures of concepts and by logical laws). An analytic truth is informal if other statements of the same form are false. Thus, *any case of knowledge is a case of belief* is an analytic truth, since its negation is incoherent, and it's not formally true since it has the same form as *any case of knowledge is a case of happiness*, which is false. See "analytic" and "entailment."

Instance of an open-sentence*:* Given some open-sentence O, if some actual sentence S is what results when the free variables in O are replaced with constants, S is an *instance* of O. So "Smith is tall" is an instance of ‹x is tall›, and "Smith is tall" is also an instance of ‹Smith has property phi.› See "open-sentence" and "quasi-quotation."

Instance of a property*:* If something has a property, then it is an *instance* of that property. We are all *instances* of the property of being human. Instances of properties are not identical with those properties themselves. I am not identical with the property of being human since my fellow humans are not instances of me. See "instantiate" and "property."

Instantiate*:* Anything that has a given property is an instance of it; and anything that is an instance of a given property *instantiates* it. You are human; so you instantiate the property of being human. See "property" and "instance."

Intensional vs. extensional definition*:* If a class has infinitely many members, then it must be defined *intensionally.* An *intensional definition* of a class is one that picks out the members of that class by citing some property

had by all and only those members. So the class of even numbers is *intensionally defined* as the class that contains every number that is divisible by two and that doesn't contain anything else.

If a class has only finitely many members, then it is possible to give an *enumerative* definition of it. Consider the class of even numbers greater than zero but less than 10. It can be enumeratively defined as the class that contains 2, 4, 6, and 8.

Technically, enumerative definitions are intensional definitions. If you identify a class by listing its members, you *are* identifying a property that they and they alone have. For example, if you identity C as the class containing 2, 4, 6, and 8, you are identifying C as the class containing all and only those objects x such that x have the property of being identical with 2 or 4 or 6 or 8. See "intension vs. extension."

Intension vs. extension*:* Intensions are to predicates what senses are to referring terms, and extensions are to predicates what referents are to referring terms. So if Beethoven's nine symphonies are Smith's nine favorite pieces of music, then "symphony composed by Beethoven" has the same extension as "one of Smith's top nine favorite pieces of music," meaning that, for any object x, ‹x is one of Smith's top nine pieces of music› is true iff ‹x is a symphony composed by Beethoven.›

But obviously those sentences have different intensions. After all ‹x is a symphony composed by Beethoven› entails ‹x is a symphony›, whereas ‹x is one of Smith's top nine favorite pieces of music› does not entail that. See "predicate" and "quasi-quotation marks."

Interpretation*:* As it is used in everyday life, an "interpretation" of a given body of data is a hypothesis as to what is meant by that data. As it used in philosophy, the word "interpretation" sometimes bears this meaning; but it sometimes bears a very different one. An "interpretation," in the distinctly philosophical sense of the word, is an *assignment* of meaning to an expression that contains undefined or only partially defined constituents. (In what follows, the words "interpretation" and "interpret" (etc.) will be used only in the second, strictly philosophical sense.)

It is statement-forms, not statements proper, that are interpreted. (A statement-form is an ***open-sentence.***) The reason is that anything with an undefined constituent *ipso facto* fails to say anything true or false. To

interpret an open-sentence is to assign constants to the variables. So one way to interpret "x is even and y is odd" would be to assign the numbers two and three to "x" and "y," respectively. That interpretation of that statement-form *validates* it, meaning that the resulting sentence is correct. By assigning three and two to "x" and "y" respectively, we've produced an interpretation of it that fails to validate it. And by assigning Barack Obama to "x'" and the property of being the U.S. President in 2009 to "phi," I've produced an interpretation of "x is phi" that validates it.

Oftentimes, expressions that *seem* not to contain free variables do. For example, "Jim is a nice guy" is so vague, that until a specific meaning has been assigned to "nice," nothing has really been said. Where expressions of this sort are concerned, context tends to supply the missing interpretations, that being why we don't see "nice" as being a variable, like "phi." Also, with "nice," there is an understood, and quite restricted, range of possible interpretations, whereas with "phi," that range is likely to be more open-ended. Thus, we're quick to see "phi" for the variable that it is, but not so quick to do so with "nice."

"Just in case"*:* See "iff."

Leibniz's Law*:* "According to *Leibniz's Law*, if x and y are numerically identical, then the one has any property had by the other. So if Ben Franklin is smart, so is the inventor of bifocals."

Logical form*:* The logical form of a sentence S is what is represented by a sentence S* that makes it clear what S entails and what entails S. Consider the sentence:

(a) "nothing snores."

Given its grammar, (a) seems to license the inference that there exists some strange entity—the nothing, or some such—that snores. But (a) doesn't warrant this inference. What (a) means is perspicuously represented by:

(b) "the property of snoring is un-instantiated."

Unlike (a), (b) isn't misleading as to what can be inferred from it. But (a)

and (b) encode the very same proposition. Since (a) and (b) satisfy these conditions, (b) is a representation of (a)'s logical form. If the relationship borne by a sentence S_1 to another sentence S_2 parallels that borne by (a) to (b), S_2 gives S_1's logical form. See "entailment" and "meaning vs. entailment-relations."[503] See "logically perfect language."

Logically perfect language*:* Let L be a language having the following property. Given any sentence S belonging to L, S's logical and grammatical forms coincide. In that case, L is a *logically perfect language.*

There can be no such language. There could exist such a language only if a sentence's entailment-relations could always be read off of its grammatical surface. This would be possible only if all entailment were *formal* entailment. But entailment is not, in general, formal entailment. Formal entailment is parasitic on informal entailment, and formal analytic truth is parasitic on informal analytic truth. The reasons for this are given in Chapters 1 and 18.

Main connective*:* See "connective."

Materialism*:* The doctrine that mental entities (e.g., beliefs) are identical with physical entities (e.g., brain states).

Meanings vs. entailment relations*:* For two sentences to have the same meanings, it is *necessary* that they be equivalent (i.e., that they be such that the one is true if and only if the other is true). So if P entails something that Q does not entail, then P and Q don't have the same meaning. For example: "Jim is an avid reader" entails that Jim can read, whereas "Jim is a lousy soccer player" does not entail that. Therefore, those two sentences have different meanings.

But even though, in order for two sentences to be equivalent, it is *necessary* that the one entail any sentence entailed by the other, that isn't *sufficient* for sameness of meaning. For any x, ‹x is a unique even prime› is equivalent with ‹x is the unique number n such that n = 2 if there is no completeness proof for arithmetic and n = 3 otherwise.› Neither sentence entails anything not also entailed by the other. But those sentences obviously differ in meaning.

For two sentences to have the same meaning, it is necessary, not only that

they entail the same things, but that they do so *in the same way*. Even though, for any x, ‹x is a unique even prime› entails the same things as ‹x is the unique number n such that n = 2 if there is no completeness proof for arithmetic and n = 3 otherwise›, they entail those same things in different ways. For example, the *way* that ‹x is even› is derived from the first is very simple, whereas the way it is derived from the second is very complex.

Modal operator*:* A modal operator is an ***operator*** (i.e., a device that, when coupled with a sentence or ordered n-tuple of sentences, yields a new sentence) that describes the *modal status* of that proposition, meaning that it says of that proposition that it is can be true, must be true, can be false, or must be false. Thus, “it is necessary” is a modal operator, since, when placed in front of “that
1 + 1 = 2,” yields:

(i) “it is necessarily the case that 1 + 1 = 2,”

which describes the proposition meant by “1 + 1 = 2” as necessarily true. And “it is possible” is a modal operator, since, when placed in front of “that Smith is in France,” yields:

(ii) “it is possible that Smith is in France,”

which describes the proposition meant by “Smith is in France” as potentially true.

In (i), “that 1 + 1 = 2” is said to be *governed* by “it is necessarily the case,” meaning only that it denotes the proposition to which the property of being necessarily true is being attributed. In (iv), “that law school will be intolerably boring” is governed by “Laura fears” since it denotes the proposition to which the property of being feared by Laura is being attributed.

Anything within an expression governed by an operator is said to fall within the *scope* of that operator. In (ii), “France” falls within the scope of “it is possible.” In (iii), “two” falls within the scope of “Jerry believes that.” See “operator” and “epistemic operator.”

Model*:* To “model” a collection of data is to produce a hypothesis that

accounts for it.

Model (second meaning)*:* A model of a set of open-sentences is an assignment of meanings to the undefined, or partially defined, expressions occurring in a set of open-sentences. Consider the following open-sentences:

(1) There is nothing that bears R to N.
(2) Anything x bears R to one, and only one, thing y.
(3) If x bears R to y, and y bears R to z, then x bears R to z.
(4) If x bears R to y, y doesn't bear R to x.

Whether (1)–(3) come out true depends on what we take "N" and "R" to stand for, and also on what we take "anything" and "nothing" to mean. Do we mean *anything at all?* Or do we mean anything *falling* into some specific class of objects (e.g., the class of numbers or of aquatic mammals)? The same question *mutatis mutandis* arises in connection with "nothing." (If we take "anything" to mean "any whole number," then the *universe of discourse* is the class of whole numbers. If we take it to mean "any penguin," then the universe of discourse is the class of penguins.)

To answer these questions is to provide an "interpretation" of these expressions. An "interpretation" of a set of open-sentences is thus an assignment of definite meanings to the undefined or partially defined expressions occurring in it. If, given a particular interpretation X (1)–(3) come out true, X is a "model" of (1)–(3). If, given X, (1)–(3) comes out false, then X doesn't model them.

Here's one interpretation of (1)–(3). Take the universe of discourse to be the whole numbers (zero, one two, etc.). N is zero, and R is the relation of being the immediate predecessor of. Relative to this interpretation, (1)–(3) come out true. This interpretation is therefore a *model* of (1)–(3).

Here's a different interpretation of (1)–(3).Take the universe of discourse to be the class of people. Take N to be Dick Cheney, and take R to be the relationship that a child bears to either one of its parents. This interpretation fails to model (1)–(3). There are several reasons for this. Thus, interpreted, (1) says Dick Cheney isn't a parent, which is false; (2) says that nobody has more than one parent, which is false; and (3) says that grandparents are immediate parents, which is false.

Molecular proposition*:* A proposition that has another proposition as a proper part. For example, *if grass is green, then something is green* is a molecular proposition, since it has two propositions as proper parts of itself, namely, *grass is green* and *something is green.* Other examples of molecular propositions are *Bob is short and Mary is tall; Mary is smarter but Sally is smarter than Mary.*

Some propositions that are in fact molecular don't *appear* to be at first. For example, *John wants to catch a fish* is a molecular proposition, since it's the same as *John wants it to be the case that John catches a fish*, and *John catches a fish* is a proposition. See "atomic proposition."

Molecular sentence*:* A sentence that has another sentence as a proper part (e.g., "grass is green and snow is white"). Some sentences are molecular that don't initially appear to be. For example, "John wants to catch a fish" is molecular since it's really an abbreviation for: "John wants it to be the case that John catches a fish."

Many of the expressions we're about to define (e.g., "quantified generalization") are used in connection with propositions *and* with sentences.

Mutatis mutandis*:* This means *provided that the relevant changes are made.* Suppose that you and I have different employers and also that your boss hates my boss with such vitriol that he will reward anyone who harms my boss in any way. In that case, the statement: "I'll get fired if I punch my boss, and you'll get fired if you do the very same thing" is false. Supposing that I punch my boss, your doing the very *same* thing would consist in your punching *my* boss, which wouldn't get you fired. But if you were to punch *your* boss, you would get fired. Thus, the right statement is: "I'll get fired if I punch my boss, and you'll get fired if you do the very same thing *mutatis mutandis.*"

Necessary*:* A proposition is necessary if there is no coherently conceivable circumstance where it is false. Example of such a proposition: 1 + 1 = 2. See "contingent," "possible," and "proposition."

Negation*:* The negation of P is not-P. So the negation of a proposition is the statement that says that P is false. (Ockham uses the obsolete term "contradictory opposite" instead of "negation.")

Necessity, sufficiency*:* "If P then Q" means that there's no way that P can be true *unless* Q is true. This, in its turn, means that the truth of Q is *necessary* for that of P and, also, that the truth of P is *sufficient* for the truth of Q. See "entails."

Necessity, sufficiency (revisited): "P entails Q" means the same thing as "Q is necessary for P," which in turn means the same thing as "P is sufficient for Q." See "entails."

Normative proposition*:* A proposition is *normative* if it says how things *should be*. *Killing is wrong* is a *normative* proposition, since it says that killing *shouldn't* occur. See "proposition" and "descriptive proposition."

Numerically identical*:* x and y are numerically identical if they are the very same thing. So Benjamin Franklin is numerically identical with the person who invented bifocals. If you punch Ben Franklin, you are also punching the inventor of bifocals and *vice versa.*

Observable characteristic*:* Any characteristic of a thing that one can see or otherwise sense-perceive. One can *see* that a thing is red. One can *feel* that a thing is solid. Thus, redness and solidity are, in at least some instances, observable characteristics. One cannot *see* that a thing consists of such and such micro-particles. Thus, the property of being composed of such things is not an observable one.

It's obvious that whether a given characteristic is observable or not is typically context-dependent. A thing's being red or solid might be as theoretical a fact as its consisting of molecules of a certain kind. To creatures of our intelligence level, but of one-trillionth our size, the fact that my desk is solid could be known only through elaborate inferences; and its being solid would not, in that context, be something that could be observed.

Still, the distinction between observable and unobservable characteristics is a good one. But, when saying that x's having phi is (un)observable, one must realize that what one is saying is elliptical for a relational statement of the form: given the nature of physical relationship to x, x's having phi is, with respect to us, an (un)observable fact.

Occurrence (as in occurrence of a linguistic expression)*:* *An occurrence of an expression is a token of it. The expression "snow" occurs three times (i.e., there are three occurrences of it) in the following sentence: "Jim likes snow and I love snow but Mary hates snow." See "expression-token."*

Open-sentences*:* Consider the sentence "Bill Gates is wealthy." That sentence is true. Replace "Bill Gates" with a variable—with an "x." The result is ‹x is wealthy›, which is neither true nor false. ‹x is wealthy› is an "open-sentence." (The characters flanking that expression are known as ***quasi-quotation marks.*** They may, for the time being, be treated as quotation-marks.)

In general, an open-sentence is an expression that contains a free variable and is therefore neither true nor false.

A *free variable* is one for which no reference is supplied either by the context or by preceding material. So if, out of the blue, I say ‹x is even›, my utterance contains a free variable, and is therefore neither true nor false.

A *bound variable* is a variable that isn't free. Thus, a bound variable is one that is assigned a referent by the context or by preceding material. Consider the sentence:

(LU) "*For any human being x, if x isn't loved by anyone, x isn't happy.*"

LU says that, *given* a human being, if *that* human being isn't loved, then *that same human being* isn't happy. Thus, the italicized part assigns a referent to the occurrences of the variable (the "x"). The occurrences of "x" in the underlined part of LU are *bound.*

An open-sentence isn't true or false. But replacing the free variables in an open-sentence may result in a true sentence. If an open-sentence is such that replacing the variables in it sometimes/always/never results in a true sentence, that open-sentence is said to be "true for some/all/no values of its variables."

Thus, "x is an even number" is true for some values of its variables (since "two is an even number" is true but "three is an even number is false"); and "x is identical with x" is true for all values of its variables; and "x is not identical with x" is false for all values of its variables.

Commentary: Logicians generally say that their discipline studies relations holding among sentences. This isn't true. It studies relations holding among the sentence-schemata. "Grass is green and show is white" entails "snow is white." The logician is interested in this fact *only* to the extent that it alerts him to the corresponding generalization, viz. for *any* sentences P and Q, ‹P and Q› entails ‹P. › But ‹P and Q› and ‹P› are not sentences; they're open-sentences; they're sentence-like entities that contain free variables and, unlike real sentences, are therefore neither true nor false.

Operator*:* An "operator" is any expression that, when joined with a sentence, forms a new sentence. So "it is possible" is an operator, since, when conjoined with "Smith is in France," a new sentence is formed, namely "it is possible that Smith is in France." And "George believes" is also an operator for the same reason.

There are many kinds of operators; but two kinds are of special importance: modal operators and epistemic operators. See "predicates" and "quasi-quotation marks."

"Other things being equal": See "*ceteris paribus.*"

Perception*: See "sense-perception."*

Performative utterance*:* When, in the context of a marriage ceremony, the priest says "I now pronounce you man and wife," the priest is not (so it would seem, and so we'll momentarily assume) reporting an existing fact. (I'll explain the raison d'être for the hedge in a moment.) He is not affirming the already existing truth that he is now in the process of forging a marital bond between the two individuals before him. In uttering those words, he is *creating* a fact, not reporting a fact.

A sentence-utterance functioning in this way is known as a "performative" utterance. Other examples of Performatives are utterances of "I christen this ship the Mel Gibson," "I hereby challenge you to a duel," and "I promise to pay you back with interest if you loan me the money."

The concept of a performative was first clearly identified by J.L. Austin in his (1955) lectures at Harvard, which were published in 1976 under the title *How to do things with Words.* Austin's points are anticipated, albeit obscurely, by Wittgenstein (1958).

Austin said, very plausibly, that performative utterances are neither true nor false. In saying "I now pronounce you man and wife," the priest isn't describing an existing fact, and his words are therefore neither true nor false —or so it would seem.

But in his (2004) book *Philosophy in the 20th Century* (Volume II), Scott Soames brilliantly shows that this is false. Consider the statement:

(S) "Supposing for argument's sake that I now pronounce you man and wife, it is incumbent on you to love and cherish each other."

(S) is synonymous with:

(S*) Supposing it true that I now pronounce you man and wife, it follows that it's true that it is incumbent on you to love and cherish each other."

S is meaningful; in fact, it's correct (or at least could be, given certain reasonable assumptions about the individuals in question). Let O be the occurrence in S of "I now pronounce you man and wife." If O cannot be meaningfully supposed true, S is meaningless, and therefore neither true nor false. Bearing this in mind, let O* be some occurrence of "I now pronounce you man and wife" that is occurring on its own. O and O* obviously have the same meaning. Since O is capable of being meaningfully supposed true, the same is true of O*, and Austin is therefore wrong to say otherwise.

An argument similar to the one just given shows that a celebrated contention of Wittgenstein's is false. Wittgenstein (1958) said that, when one says "I am sad"—when, in general, one makes any statements about one's state of mind—one isn't really *saying* anything. Saying "I am sad," Wittgenstein says, is no more a *bona fide*speech act than is grimacing or moping about. In saying this, Wittgenstein was trying to undermine the view that sentences report facts; he was trying to vindicate his provocative—but, when scrutinized, quite hollow—contention that "words are deeds." (This contention is one to which Austin gave some substance in his theory of performative utterances.)

We can use an analogue of Soames' argument to refute this contention of Wittgenstein's. If, as Wittgenstein alleges, a stand-alone occurrence of "I am sad" is neither true nor false, as it's merely a glorified frown, then it's counterpart in an utterance of

(S#) “if it’s supposed true that I am sad, then I am in a mood that is, in at least some respects, less than optimal”

is equally meaningless, an immediate corollary being that any utterance of S# is meaningless. But such an utterance is quite meaningful; in fact, such utterances are correct.

Also, if you say “I am sad,” you are obviously reporting a fact. That’s why your utterance could be false—you could be lying after all. Thus, utterances of “I am sad,” and first-person reports generally, can be meaningful, and usually are, contrary to what Wittgenstein is alleging.

Personal stratum of cognition*:* The totality of mental events within one’s own mind that could in principle fall within the scope of introspective awareness. Mental events in your mind mediate between sensory input and cognitive output. You could not become *introspectively* aware of these. Nor are you aware of the cognitive events in your own mind that mediate language-learning and comprehension. So far as you can become aware of them, it is in the way that you can be aware of sub-atomic particles, and not in the direct, introspective way in which you are aware of your own sensations. No amount of psychoanalytic intervention could make you introspectively aware of them, and they are therefore “deeply” unconscious, to use Chomsky’s term. But, given psychoanalytic intervention, you could become introspectively aware of the unconscious events and conditions posited by Freud. Thus, the Freudian unconscious falls within the scope of the *personal* stratum of cognition, as does everything that, without such intervention, is accessible to introspection. The cognitive processes mediating language-comprehension, perception, etc., are not done by *you*, even though they happen *in* your mind. There is thus a sense in which they are impersonal (or subpersonal). They constitute a framework underlying the events constitutive of the personal stratum of cognition and, for that reason, are accessible to it.

So far as I know, Chomsky (1959) deserves credit for being the first to delineate the concept of subpersonal thought, and also for being the first to put the concept of subpersonal mentation to scientific use. In doing the former, he opened up new vistas, not just in psychology, but in epistemology and philosophy generally. There are allusions to the subpersonal in Chapter

VII of Freuds (1901) work *The Interpretation of Dreams.* The main contention of Kant's (1789) work, *The Critique of Pure Reason*, presupposes its existence. Perception and thought, Kant contends, result from a convergence of (i) disturbances of our sensory surfaces and (ii) concepts that are hardwired into us and constitute our innate cognitive endowment. Kant makes it very clear that it is not until *after* these concepts have processed these disturbances that there occurs any mental activity that could be attributed to a *person*. So Kant clearly believes that personal mental activity presupposes the existence of pre-personal—or, as we might also put it, subpersonal—mental activity. Chomsky often cites Kant as one of his forbears.

Perspicuous/perspicuity*:* A sentence is "perspicuous" if it is clear *in the sense that* its inferential structure (what it entails and what entails it) can be read off of its grammatical structure. A perspicuous, set-theoretic translation of "2 + 2 = 4" would be totally unclear to most people, even though "2 + 2 = 4" is very clear (in the psychological sense). That's why "clear" and "perspicuous" aren't interchangeable. Perspicuity is a logical property; clarity is a psychological property. Perspicuity is to some extent a contextual notion. Depending on what one's background assumptions are, one sentence may or may not be more perspicuous than some other synonymous sentence. But, other things being equal, the more syntactic structure a sentence has, the more perspicuous it is. The reasons for this are given in Chapter 18, Section 4.1.)

Possible*:* A proposition is possible if there is some coherently conceivable circumstance where it is true. Example of such a proposition: Mary owns a BMW. See "contingent," "necessary," and "proposition."

Predicate*:* In the sentence, "Smith is tall," the *predicate* is "tall." A "predicate" is generally said to be an expression that refers to a property. But this isn't a good definition. The expression "the property of being tall" is not a predicate; it's actually a singular term. But "tall" *is* a predicate. So is "smart."

In general, a predicate is an expression that, when joined in a grammatically acceptable manner with a singular term, yields a sentence that attributes a property to the individual referred to by that term. "Smart" is a

predicate because, when joined (in a grammatically acceptable way) with "Sally," the result is a sentence that attributes smartness to Sally, viz. "Sally is smart."

Predicates may be identified with open-sentences. See "intensions vs. extensions" and "open-sentences."

Primitive expression*:* Synonymous with "semantically simple expression." A primitive expression is one that doesn't consist of other expressions. Examples of such expressions are "red" or "sweet."

Occurrences of primitive expressions are not as easy to identify as one might think. Let T be an utterance of "snow is white." One would think that, in T, the occurrence of "snow" is semantically simple.

Chomsky would say that it contains a phonetically unrealized case-marker. Chomsky's reasons for taking this view are extremely compelling. See Chomsky (1965, 1998).

Probability*:* This word is "ambiguous." It has (at least) two meanings. First, there is the statistical meaning of "probability." If there are 10 balls in the urn, only one of which is white, the statement "x is a white ball in the urn" has a statistical probability of 1/10.

Then there is explanatory probability. P makes Q probable in the *explanatory* sense if, given what it is that is already believed, there are more causal anomalies if P is true and Q is false than there are if, other things being equal, P is true and Q is true.

It is argued in Chapter 11 that explanatory probability cannot be reduced to, or understood in terms of, statistical probability.

Property*:* A property is a characteristic. Other words for "property" are "attribute" and "feature." This is the same as saying that a property is anything of which there can be instances. Put another way, a property is anything that can be *had*—"had" in the sense in which one has a characteristic, not in the sense in which one has a car. So the "had" in question is that of *attribution*, not of *possession.* A property is anything that can be meaningfully attributed to something.

Commentary: some exotic properties: Consider the number two. There are *instances* of this property. Any pair of objects is such an instance. The

number two can thus be thought of as a property had by all pairs of objects and by nothing else.

Anything of which there are *instances* is a property. (But there are some properties of which there are no instances; e.g., the property of being a perfectly moral human. But that property is composed entirely of things of which there are instances. So it's possible—though I haven't personally verified it—that every property *either* has instances *or* is composed of properties that have instances. Question: are there properties that counterexample this conjecture?)

The *numeral* "2"—that is, the *sign* for the number two—is *also* a universal. But it's distinct from the number two itself. The numeral "2" is, for reasons that we discussed a little while ago, a property of certain ink marks and bursts of sounds. (Those ink marks, etc., are *tokens* of that numeral.) But the number two is a property of *pairs* of objects. "2" is no more identical with the number two than the name "John-Michael" is identical with the person John-Michael.

Works of music are properties. *Instances* of those properties are what we refer to as "performances." Consider the first movement of Beethoven's "Moonlight Sonata." I played it a few minutes ago. But that work of music didn't cease to exist when I stopped performing it. Of course, it's likely that other people are performing it. But it wouldn't cease to exist if, for a 10-minute period, *nobody* were to perform it. The "Moonlight

Sonata" isn't identical with any *particular* stream of sound. Therefore, it isn't identical with any stream of sound at all. The "Moonlight Sonata" is a *property* of certain streams of sound—of streams of sound that have certain melodic and harmonic properties. When somebody plays that Sonata on the piano, that person is producing a stream of sound that has the right properties and is therefore an instance—or, as we would usually say, a performance—of that sonata. (Other works of art, e.g., paintings, appear to be concrete objects. The Mona Lisa is a concrete object; it could be destroyed. Works of visual art seem, in general, to be spatiotemporal and thus to be property-instances, as opposed to properties. Whether works of literature are properties or property-instances is a delicate question that will be set aside.)

Propositions*:* The things that are *meant* by true or false utterances or inscriptions.

Commentary: Propositions aren't sentences; propositions are sentence-meanings, not sentences per se.

Propositions are not spatiotemporal. The proposition *snow is white* does not have a location in space or time, even though particular *instances* of white snow obviously do have such a location.

Propositions are not psychological entities; they are mind-independent entities.

Believing a proposition is something mental; but the thing believed is not. More generally, *grasping* a proposition is something mental; but the thing grasped is not.

Let's suppose that:

(1) "Smith believes that the inventor of bifocals is 6-feet tall'

is true.

What is the *object* of Smith's belief? What is the *thing believed in*—the thing that, in having that belief, he believes to be true?

It's tempting to say that it's something *mental*—that it's some mental event or psychological condition. But this is wrong. Smith's *having* that belief is indeed a psychological condition. But the thing *believed* is not. The believ*ing* is mental; the believ*ed* is not. In other words, the thing that, by virtue of having that belief, Smith regards as true isn't a mental or psychological event or condition (or whatnot) at all. Here's why.

Whatever it is, that thing is something that *others* grasp. (And, what follows, it is also something people besides Smith *could* grasp, even if they don't.) Those who agree with Smith (about the height of the inventor of bifocals) believe it true; those who disagree believe it false. But they all grasp that thing.

Supposing that any given one of those people were to perish, and that all his or her mental contents were therefore to vanish, these other people could (and probably would) continue to grasp that thing; they could continue to believe it true or false or probable or whatnot. If Smith and I both believe that the inventor of bifocals is 6-feet tall, Smith's dying won't prevent me from continuing to believe it or, therefore, from continuing to *be able* to believe it. Smith's death, though tragic, is irrelevant.

By the same logic, appropriately generalized, given *any* two people who believe or otherwise grasp *any* proposition, the existence of the one person's

mind and mental states is independent of the other person's believing (or otherwise grasping) that proposition. *Any* one person's mental states and events can cease to exist without in any way jeopardizing anyone's ability to believe or disbelieve that the inventor of bifocals is 6-feet tall.

Thus, the thing that Smith believes—the *object* of his belief—is something that continues to exist even if any given person's mental events and conditions cease to exist. Therefore, that thing isn't identical with anyone's mental states or, needless to say, with anyone's mind in its entirety. Nor is it identical with any *thing's* mental states. Supposing that non-humans—animals orMartians or whatever—have beliefs concerning the height of the inventor of bifocals, everything we just said holds of those beliefs. Since that thing isn't identical with the mind or mental states of any entity, it can't be identical with *any* mental entity. To deny this would be to affirm the absurdity that there is a mental thing which isn't any particular mental thing.

What, more exactly, is the nature of this thing? What, in general, are propositions? (In other words, what is the nature of the things that are the objects of belief and disbelief? We don't have to worry about that right now; it's irrelevant to what we're about to say. But it is thoroughly answered in Chapter 3.)

Additional commentary: Propositions have the distinctive and crucially important property of being *non-derivatively true or false.* What does this mean? Propositions are obviously true or false. (It's true that Finland is in Europe; it's false that it's an island in the Pacific Ocean.)

To be sure, things other than propositions are true or false (e.g., sentences and beliefs). The sentence "Finland is in Europe" is true. But that sentence is true *only by virtue of its association with the corresponding proposition.* That sentence is true only because it expresses a true proposition. So it's really the proposition that deserves to be described, without qualification, as "true"; and when we say of the sentence that it's "true," we're using the word "true" to pick out a related, but distinct, property; we're using it to pick out the property of *encoding* something that's true.

And these, it must be stressed, are very different properties. It's one thing to have the property of being true. (The proposition *Finland is in Europe* has that property. It's quite another to have the property of *encoding* something true.) The sentence "Finland is in Europe" has *that* property. But the corresponding proposition does *not* have it, since propositions, unlike sentences, don't encode anything.

What we just said about sentences is true of beliefs. My belief that Finland's in Europe is "true." But it's true *only because of its association with a true proposition.* In other words, it's "true" in the sense that it has for its content something that is true. Put another way, the thing which that belief is a belief *in* is true, and the belief is "true" only in the second-class sense that it has this association with this other thing—this other being something that, unlike the sentence, is true in a first-class, non-derivative sense.

Propositions are thus "non-derivatively true," whereas beliefs and sentences are "derivatively true." For this reason, propositions are sometimes defined as "non-derivative bearers of truth and falsity."

We thus have three different definitions of "proposition": (i) thing meant by a sentence; (ii) thing towards which one can have an attitude (e.g., belief, disbelief, fear, hope); (iii) non-derivative bearer of truth or falsity. While no one of these definitions is 100% accurate each is accurate to a high degree of approximation; and while no two of them are precisely equivalent (i.e., while no two of them pick out precisely the same class of entities), any two of them are almost equivalent (i.e., the classes picked out by any two of them overlap to a very high degree).

Propositional attitudes*:* Propositions are the things we affirm. They are also the things towards which belief, doubt, fear, and all other *attitudes* are directed. One can't just believe; one must believe that snow is white or that grass is green. One can't just fear; one must fear *that Mary will come home before tomorrow*, or some such.

We'll see in a moment that there are things other than propositions that we believe, doubt, etc. But, though it's therefore not entirely accurate to say that propositions are *the* things we believe, doubt, etc., it's a good point of departure; and, for the time being, we'll assume it's truth.

In believing that Sally will come home tomorrow, one has a certain propositional *attitude.* In fearing it, one has a different propositional attitude. The proposition is the same, but the attitude has changed. In wondering whether Sally will come home, one has yet another attitude towards that same proposition. Belief, doubt, wonder, etc., are thus "propositional attitudes."

Knowing what is going on in somebody's mind consists largely, though by no means entirely, in knowing what that person's propositional attitudes are. And to know this, one must know which propositions are the objects of that person's mentation; and one must also know, for any given one of those

propositions, which attitude that person has towards it. To know what is going in Larry's mind, it isn't enough to know that the *object* of his thoughts is the proposition *Sally will come tomorrow;* one must also know what sort of attitude he has towards that proposition. Does he believe it? Hope that it's true? Fear that it's true? And one must have comparable knowledge of every other proposition that is the object of Larry's thought. If one doesn't have such knowledge, one is ignorant of an extremely important dimension of Larry's existence; and if one does have it, one knows much about Larry—but not everything.

Why not everything? Because not everything mental is a propositional attitude. Nothing that isn't *content-bearing* is a propositional attitude. (To say of something mental that it's "content-bearing" is to say that it bears some kind of message. So my current visual perception is content-bearing, since it tells me that there's a computer screen in front of me. My headache isn't content-bearing, since it doesn't bear any message; it's merely felt.) And many mental entities that are content-bearing aren't propositional attitudes, the reason being that not all contents are propositions. See "propositions."

Qua*:* Synonym of "in virtue of." Larry eats a fish. He's allergic to all food-products that contain protein X. This particular fish (though not others) contains X. So Larry falls ill. Jerry eats a fish; he isn't allergic to fish, but the fish he eats is rotten. So he falls ill. What made Larry sick wasn't his eating a fish; it was his eating something that had X in it. That thing happened to be a fish, but that wasn't what was operative. Thus, it wasn't in virtue of his eating a fish, but in virtue of his eating something containing X, that Larry became sick. Put another way, it was his eating a fish *qua* thing that contained X, not a fish *qua* fish, that made Larry sick. Similarly, it was Jerry's eating a fish *qua* rotten thing, not a fish *qua* fish, that made him sick.

Qualitatively identical*:* x and y are *qualitatively* identical if they resemble each other to a maximally high degree but may nonetheless be *distinct individuals.* So if you have an identical twin—a twin who looks, thinks, talks, acts (and so on) just like you—then you are qualitatively identical with that person. But you are not *numerically* identical with that person, since you are not literally the very same individual as that other person.

Quantified generalization*:* Any sentence or proposition that says of some property how many instances it has. Thus, *nothing snores* is a quantified generalization since it says of the property of being a snorer that it has no instances. For the same reason *mutatis mutandis*, each of the following is a quantified generalization: *everything snores*, *something snores*, *three things snore*, and *most things snore.* See "quantifier."

Quantifier*:* An expression that, when coupled with a predicate yields a sentence saying how many things of a given kind have that property. A predicate may *roughly* be said to be an expression denoting a property (e.g., "snore," "is tall"). (Why only roughly? See below.) "All people" is a quantifier, since, when coupled with "snore," which denotes the property of snoring, yields "all people snore," which says how many things of a certain kind have the property of snoring. "No fish" is a quantifier since, given the expression snore, yields the sentence "no fish read Shakespeare," which says how many things of a given kind have the property of reading Shakespeare.

In logic and the philosophy of language, one often uses artificial analogues of expressions like "all," "some," "none," etc. In logic, the sentence "all people snore" is represented as "for any x, if x is a person, then x snores." And "some people snore" becomes "for some x, x is a person and x snores." The reason why artificial quantifiers (e.g., "for any x," "for some x") are used in lieu of their natural-language counterparts ("all," "some") is that, for the reasons discussed in the previous chapter, the logical form of quantified statements is concealed unless these replacements are made. For example, "no person snores" has the same grammatical form as "John snores," which misleadingly suggests that, like the latter, the former attributes the property of snoring to some entity. But "nothing snores" does *not* do that; that's the very opposite of what it does. What "no person snores" says is this: "for any x, if x is a person, then x does not snore." The latter sentence clearly *doesn't* attribute the property of snoring to anyone or anything, and is, in this respect, more perspicuous than its natural language counterpart.

The reason it's only roughly correct to say that "predicates" denote properties is that "the property of snoring" and "the property of being tall" clearly pick out properties—but they aren't predicates. See "predicate."

Commentary: All quantifiers are considered to be connectives. Thus "for all x," "for no x," etc. are connectives. Here is a very rough explanation of

the rationale for this practice. Consider the sentence “all people snore.” This can be thought of as saying that each possible sentence of the form “x snores,” where x is some person, is true. So it can be thought of as saying that “John snores,” “Sally snores,” are all true. Thus, “for all x,” can be thought of as an expression that, when given “Sally snores,” “John snores,” etc., as inputs, yields another sentence, viz. “it is true that John snores, Sally snores [etc.].” I must point out that this explanation is very approximate. But the technicalities that would have to be gone through to make it completely accurate are not, at this stage, very important.

Given that all quantifiers are connectives, it follows that all quantified generalizations are considered to be molecular propositions. See “quantified generalization.”

Quasi-quotation marks*:* In logic and the philosophy of language we often wish to talk about whole *classes* of sentences having a given form. We might wish to say, for example, that *all* sentences of the same form as “grass is green or it is not the case that grass is green” are true, that *all* sentences of the same form as “grass is green and it is not the case that grass is green” are false, that *all* things have the property of being self-identical, that *no* thing has the property of a being a square circle.

But it’s hard to do this without producing a nonsense-statement. Consider the following statement:

(1) For any proposition P, the expression “P or not-P” is true.

(1) seems okay at first. But it isn’t. It’s false. This is because “P or not-P” isn’t a sentence. If, out of the blue, I say to you “x is larger than four,” I haven’t said anything, since “x” is a variable, not a numeral. (If I don’t say it out of the blue, and it’s understood that “x” has a certain referent, then it may be true.) Similarly, so “P or not-P” is no meaningful statement can contain a free variable. If I say to you, “P or not-P,” I haven’t said anything, the reason being that “P” is a variable. “P or not-P” isn’t true or false, and it’s no more capable of being true or false than an out-of-the-blue utterance of “x is larger four.” Given that “P or not-P” is never true, it follows that (1) is false. After all, (1) says that “P or not-P” is *always* true—when, in fact, it’s *never* true (or false).

But what (1) is *meant* to express, even though it fails to do so, is that,

given the expression “P or not-P,” the result of replacing both occurrences of “P” with a true or false sentence is itself a truth. (For example, if we replace those occurrences with “snow is white,” the result is “either snow is white or it is not the case that snow is white,” which is true.)

The sentence:

(2) For any proposition P, the expression ‹P or not-P› is true

is an abbreviation for

(2) Given the expression “P or not-P,” the result of replacing both occurrences of “P” with a true or false sentences is itself a truth.

(1) is often abbreviated into:

(3) For any P, ‹P or not-P› is true,

which, in its turn, is often abbreviated into:

(4) P or not-P.

The strange rectangular expressions in (3) are known as “quasi-quotation marks.” Putting quasi-quotation marks around an expression results in an expression that refers to everything in the original expression *besides the variables.* So ‹if x is larger than y and y is larger than z, then x is larger than z› is an expression that refers to everything in the expression in between the quasi-quotation marks *besides* the occurrences of “x,” “y,” and “z.” Quasi-quotation marks are the invention of Willard van Orman Quine (1941).

Quasi-quotation marks are extremely useful. Without them, or some other comparable device, we couldn’t express the truth expressed by the sentence:

(4) For all values of x, y, and z, the sentence ‹if x is larger than y and y is larger than z, then x is larger than z› is true.

Quasi-quotes, though seldom *overtly* present in discourse, are very often *implicitly* so. I’ve shown elsewhere that, without them, it isn’t possible to make what is meant by expressions that have “discourse-internal

referents,"—for example, the occurrences of "that person" in the sentence: "if a person is mean to everyone, that person will have no friends and that person will therefore be unhappy."

Some more illustrations: The sentence:

(5) For any individual x, the expression "x is identical with x" is true,

is false. This is because "x is identical with x" is no more true or false than an out-of-the-blue utterance of "x is even." What *is* true is the statement that:

(6) A true sentence results if the occurrences of "x" in "x is identical with x" are replaced with an expression, the same one in both cases, of an expression that picks out some object.

The standard way of abbreviating (5) is:

(7) For any x, ‹x is identical with› is true.

Quasi-quotes will occur frequently in the chapters on the philosophy of language. It's not always easy to know when they occur; and I often lapse, using quotation marks instead of them, or just omitting them. When I do this, mentally insert them—otherwise what I'm saying will be false. (I appreciate it.) See "quotation marks."

Quotation marks*:* The word "snow" has four letters. But snow doesn't have four letters. So the sentence "snow has four letters" is false. Snow is white, crystalline substance, which consists, not of letter, but of H_2O molecules. In the philosophy of language, we often wish to refer to expressions themselves—we often wish to talk, not about snow (the substance), but "snow" (the word that substance).We do this by putting *quotation marks* around the expression to which we refer. So "snow" refers to the *word* that refers to the substance. See "quasi-quotation marks."

Rationalism*:* The doctrine that not all knowledge is derived from sense-perception, the reason being that some knowledge is acquired through the analysis of concepts. Rationalists typically hold that knowledge is a prerequisite for our acquiring sensory information and also for our making

any inferences from it. Since such knowledge couldn't possibly be observation based, it must have some other basis, this other basis being "pure reason." Rationalists hold that knowledge of logical norms is non-perceptual, their argument being that sense-perception apprises us of what is, not of what ought to be (even in a narrowly logical sense). See "empiricism."

Reductio ad absurdum*:* An argument that proves its point by showing that the negation of that point, if assumed true, leads to a contradiction (a statement of the form *Q and not-Q*). Suppose that P is true, that you know this, and wish to prove it to a friend of yours who doubts it. One way of doing so is to assume the negation of P and then show that has a self-contradictory result. Your position is that there are no even primes greater than two. The negation of your position is that there are even primes greater than two. You apprise your friend of your viewpoint. He disagrees. You respond by saying: "Let's suppose you're right. In that case, there is some number n that is greater than two and is even and is prime. Being prime, n is divisible only by itself and one. Being greater than two, $n \neq 2$. And being even, n is divisible by two. Thus, n is divisible by itself and one *and* by two. Therefore it isn't divisible *only* by itself and one. So n both is, and is not, divisible *only* by itself and one. You've reduced your friend's position to absurdity.

Refer*:* The word "refer" is ambiguous. It may refer to a property of actions or to a property of expressions. (In the present work, it is almost always used to denote a property of expressions.)

In an utterance of the sentence "when I was discussing the difference between real virtue and feigned virtue, I was referring to your father," the occurrence of the word "referring" picks out a characteristic of a certain ***speech act.*** Given some object O, a given act A that is performed by so and so is a case of so and so's referring to O if, in performing A, it was so and so's intention to make a statement about O. Let's refer to this sort of reference as "reference*."

In the sentence, "the name 'Socrates' refers to a certain philosopher," the occurrence of the word "refers" picks out a property of expressions, not of actions. To say that an expression E refers to some object O is to say that, in virtue of containing E, a sentence *ipso facto* concerns O. "Socrates" refers to Socrates because, in virtue of having the form "...Socrates...," a sentence says that Socrates has...x...("Socrates is tall" says that Socrates has the property of

being a thing x such that x is tall. “Plato admires Socrates” says that Socrates has the property of being a thing x such that Plato admires x.) Let us refer to this sort of reference as “reference#.”

Some authors[504] have tried to show that reference# was identical with reference* or, at the very least, capable of being understood in terms of it. This is radically absurd.[505] Reference* presupposes reference#. Nobody could refer* to anything unless there were already meaningful expressions and, therefore, referring# terms. In Chapter 4, Sections 5.3.2–5.3.4, it is explained why this is so. (So far as I know, it was John Searle, in his classic (1966) work *Speech Acts*, who made this important point for the first time in print.)

Referent*:* the reference of an expression is what it refers to. See “refer.”

Scope*:* See “modal operator.”

Second-order desire*:* A desire about a desire (e.g., a desire to refrain from acting on one’s desire to have a cigarette).

Self-contradictoriness*:* Any statement that has the form ‹P and not-P› or that implies a statement of that form is self-contradictory. All self-contradictory statements are false. Anything that entails a self-contradictory statement is false.

Semantic externalism*:* The doctrine that what our words and sentences mean isn’t determined wholly by what our thoughts are, but also by environmental factors of which we may be unaware.

Semantic rule*:* A rule that assigns meaning to an expression. It’s a semantic rule of English that “snow” refers to a certain crystalline white substance, and “water” refers to H_2O, etc.

At least some semantic rules assign meaning to expression-types, on the basis of which tokens of those types have this or that meaning depending on the circumstances. Thus, the meaning of the word-type “I” doesn’t assign this or that individual to it. “I” per se doesn’t refer to you or me or anyone else. But “I” does have a meaning. That meaning is a rule to the effect that tokens

of it refer to the people who produced those tokens. See “expression-type.”

Sentential force*:* Some sentences are true or false (e.g., “snow is white”). But some are neither true nor false. Imperatives (i.e., orders—e.g., “get out of here”) and interrogatives (i.e., questions—e.g., “what’s a hypotenuse?) are neither true nor false. True or false sentences are known as *indicatives*, and they therefore differ in respect of *force* from both interrogatives and imperatives.

Sense vs. reference*:* “The inventor of bifocals” refers to Benjamin Franklin, and so does “the first postmaster general.” But, even though those expressions co-refer, they have different *senses.* Although Benjamin Franklin was the first postmaster general, it isn’t in *virtue* of his having this property that “the inventor of bifocals” refers to him. And although Benjamin Franklin was the inventor of bifocals, it isn’t in *virtue* of his having this property that “the first postmaster general” refers to him. It is in virtue of his being a unique first postmaster general and a unique inventor of bifocals that Benjamin Franklin is the referent of, respectively, “the first postmaster general” and “the inventor of bifocals.” The *sense* of an expression is the property that an object must have to be its referent. So supposing that O is the referent of some expression R, and that it is in virtue of O’s having property P that R refers to O, P is the “sense” of R.

Commentary: Those who hold that definite descriptions are quantifiers, as opposed to referring terms, will reject what was just said. See Chapter 6. See “definite descriptions” and “quantifier.”

Sense-perception*:* Sight, hearing, touch, taste, smell, and one’s bodily kinesthetic sense (the feeling of resistance one experiences when trying to move objects—this sensory modality involves touch but isn’t identical with it). The two distinguishing features of sense-perception are (i) that it yields knowledge that is (at the ***personal level of cognition***) non-inferential, and (ii) perceptual information varies in real-time with the states of affairs it concerns. See “personal stratum of cognition.”

Sets and classes*:* A class is simply a collection of some kind—for example, the class of numbers, the class of even numbers, the class of people,

the class of people who are taking philosophy 101 this semester, the class of smart people, the class of talented athletes.

The word "set" is sometimes used instead of "class." We'll use these terms interchangeably. (Some set-theorists, e.g., Quine (1956), do not do so.)

Commentary: The sets that are of intellectual interest are usually ones whose members have some property in common with one another that nothing else has in common with any of them. So, for example, we talk about the set of whole numbers, the set of people who have survived pancreatic cancer, the set of people who are currently philosophy professors. Let S_1, S_2, and S_3 be these three sets. Every member of S_1 has the property of being a whole number and nothing that isn't a member of S1 has that property. Thus, the property of being a whole number *generates* S1. The property of being a survivor of pancreatic cancer and the property of being a philosophy professor generate S_2 and S_3, respectively.

Technically, given *any* set, there is *some* property that all and only its members have in common. But oftentimes these properties are very contrived. Let S be the set consisting of Smith and the number two. Everything in S has the property of being identical either with Smith or with the number, and nothing outside of that set has that property. But, of course, we're usually interested in sets that are generated by explanatorily more robust properties (e.g., the property of being a number).

Additional commentary: Sets ≠ properties: Some have held that sets are with properties. Their argument: "Smith is a person" (i.e., "Smith has the property of being a person") is equivalent with "Smith is a member of the class of people," and the same thing *mutatis mutandis* holds of any statement that attributes any property to anything.

But sets are not properties, and it's easy to identify the fallacy in the foregoing argument. "Smith is a person" is *not* equivalent "Smith is a member of the class of people." Let S be the class of people currently alive, and suppose that both Smith and Jones are members of S. Remember that sets are individuated by their memberships. Thus, S couldn't possibly exist unless Jones did as well. But Smith could exist and be a person in a world where Brown and, therefore, S failed to exist. So "Smith is a person" is not equivalent with "Smith is a member of S."

Sets are modally and temporally frozen. Properties are modally and

temporally elastic. A given property may have different instances at different times and in different universes. (Some properties are modally frozen—e.g., the property of being an even number. But such properties are not representative of properties in general, and we may disregard them in what follows.) Thus, for any property phi (with the parenthetically noted exceptions) and any set S, ‹x has phi› is not equivalent with ‹x is a member of S.›

Simpliciter*:* This means *without qualification.* "There is no such thing as motion *simpliciter*—one must be moving *relative to this or that framework.*"

Singular term*:* Any terms refers to an individual (e.g. "Socrates," "Shamu"). See "referring term."

Slingshot-style arguments*:* Consider the following sentences:

(1) Mozart was a composer.
(2) The class of all things x such that ***Mozart was a composer and x is self-identical*** is identical with the class of all things x such that ***x is self-identical***.
(3) The class of all things x such that ***Bill Clinton was U.S. President in 1997 and x is self-identical*** is identical with the class of all things x such that ***x is self-identical***.
(4) Bill Clinton was U.S. President in 1997.

Here is the position of those who believe in the cogency of slingshot-style arguments:

(SS[506]) (1) is logically equivalent with (2). Therefore, (1) and (2) must co-refer, supposing that either refers to anything. (It is being assumed, very reasonably, that logically equivalent expressions cannot have different referents.)
(3) is what results when a referring term in (2) is replaced with a co-referring term. ("The class of all things x such that Mozart was a composer and x is self-identical" refers to the class of all self-identical objects, and so too does "the class of all things x such that Bill Clinton was U.S. President in 1997 and x is self-identical.") Since

intersubstituting co-referring terms preserves reference, it follows that (2) and (3) co-refer, supposing that either refers to anything.

(3) is equivalent with (4). Therefore, (3) and (4) must co-refer, supposing that either refers to anything.

Therefore, (1) and (4) must co-refer, supposing that they refer to anything.

If the occurrences in (1)–(4) "Mozart was a composer" and "Bill Clinton was U.S. President in 1997" are replaced with occurrences of *any* two true sentences, the argument that results is cogent. Thus, any two true sentences co-refer, supposing that any true sentence refers to anything. (For the same reason *mutatis mutandis*, any two false sentences co-refer, supposing that any false sentence refers to anything.)

The property of being true is the one semantically significant thing that (1) and (4) have in common. So, supposing that sentences refer to anything, (1) and (4) must refer to the property of being true, and true sentences must in general be assumed to refer to the property of being true.[507] (For the same reason *mutatis mutandis*, false sentences must refer to the property of being false.[508])

In general, a slingshot-style argument is one to the effect that, since logically equivalent expressions can't have different referents and since intersubstituting co-referring terms preserves truth-value, some class that would seem to comprise several objects (for example, the class of things that are referred by true sentences) in fact only comprises one object (for example, the property of being true).

SS is a poor argument that embodies a spurious understanding of reference, the same being true of every other slingshot-style argument. Suppose for argument's sake that "the class of all things x such that Mozart was a composer x is self-identical" is functioning as a device of reference. In that case, it's picking out the class of self-identical objects, and what (2) is saying about it is that it's identical with itself; in other words, (2), thus interpreted, is saying that:

(2*) the class of self-identical objects is identical with the class of self-identical objects.

But (2*) is not equivalent with (1). Therefore, if the definite descriptions in (1) and (2) are assumed to be referring terms—which is precisely what they *are* being assumed to be—(2) isn't equivalent with (1). For the same reason *mutatis mutandis*, (4) is not equivalent with (3) if "the class of all things x such that Bill Clinton was U.S. President in 1997 and x is self-identical" is taken to be a device of reference.

In order for (2) to be equivalent with (1), the former must be taken to say that:

(2#) there exists a class to which a thing belongs iff Mozart is a composer and that thing is self-identical; and there exists a class to which a thing belongs iff that thing is self-identical; and the first class is identical with the second class.

But if (2) is interpreted in this way, the definite description occurring in it is a quantifier, not a referring term. In which case, (3) is *not* what result when a referring term is replaced with a co-referring term. (It's what results when a non-referring term is replaced with a non-referring term.)

So SS is a failure.

Also, it's obvious that "snow is white" doesn't co-refer with "the property of being true" (or with "the *True*"). In virtue of having the form, "...the property of being true..." a sentence is about the property of being true. ("Everything Bill says has the property of being true" says of the property of being true that it is instantiated by everything that Bill says.) But in virtue of having the form "...snow is white..." a sentence is about the *proposition* that snow is white. ("If snow is white, then snow isn't green" says that the proposition that snow is white, if true, has the proposition that snow isn't green as a consequence. Of course, "if snow is white, then snow isn't green" can, like any other sentence, be interpreted as concerning the property of being true—it can be interpreted as saying that the property of being true is a characteristic of the proposition that snow isn't green if it's white. But it isn't *in virtue of* it's containing an occurrence of "snow is white" that "if snow is white, then snow isn't green" concerns that property.) So unless the concept of reference is warped beyond all recognition—unless the term "reference" is redefined to as to validate this argument—what SS establishes, supposing its premises true, is false. In any case, we've seen independently of this that not

all of it's premises are true.

But Donald Davidson took it for granted that SS is cogent.[509] And he argued that, given the (alleged) fact that SS is cogent, it follows that there is only one fact. His argument is simple. Put the expression "the fact that" in front of each of (1)–(4). The same principles that show that (1) and (4) must co-refer also shows that "the fact that Mozart was a composer" co-refers with "the fact that Bill Clinton was U.S. President in 1997)."

Davidson concluded, absurdly, that these two facts *are* identical—and that, in general, there is only one fact. Having, in his mind, established that there is only one fact, he argued that mental states are not representations. Since there is only one fact, Davidson argued, no two mental states represent different facts. On this basis, Davidson concluded that differences between mental states—between, for example, my perception of the piano in my living room and your perception of the ink-marks in front of you—are not representational differences. Why is this? Because, says Davidson, representations, supposing them to exist, are representations of facts; and, as we've just seen, there is only one fact. Therefore, Davidson concludes, my perception of the piano in my living room is *representationally* just like your perception of these ink-marks. Thus, the differences between those perceptions have nothing to do with any representational differences that there might be between them. Which means that, if there are any such differences, they are innocuous. Which, in its turn, means that *either* (i) that neither perception represents anything or (ii) that, if either perception does represent something, it is irrelevant to how that perception differs from other perceptions (or mental states of other kinds) and, therefore, to how what perception is. Which, give or take some unimportant nuances, means that neither represents anything.

Davidson's position (that no mental entity represents anything) is obviously false, and his argument for it fails for the same reason that SS fails. But Davidson's argument involves a blunder of which SS is innocent. For a mental state to be representational is not for there to be some fact that it represents. It is for that mental state to be to the effect that there exists some fact. Suppose that I hallucinate that there is a pink elephant in my living room. (Assume that, in actuality, there is no such elephant.) My hallucination is representational. It isn't like a rock. But there is no elephant x such that my hallucination represents some pink elephant x such that x is in my living room. My hallucination is representational in the sense that it is to the effect

that that there exists some pink elephant x such that x is in my living room. (See Chapter 1, Section 3.2 for further clarification.)

In his (ironically named) book *Facing the Facts*[510], Stephen Neale puts forth what he believes to be an improvement on SS.[511] Neale's argument involves the following four assumptions:

G_1: The sentence "Fa" stands for the same fact as the sentence "a=the unique thing x such that (x = a and Fx)." (In this context, "=" is to be taken to mean "is identical with.")

G_2: Any sentence can be put into subject-predicate form.

G_3: What a complex expression refers to is a function of what it's parts refer to.

G_4: Definite descriptions are referring terms. "The even prime" is an expression that picks out the number two.

Assume that for some individuals a and b, and some properties G and F, the following are true: "Fa," "Gb," and "a ≠ b"; and suppose that these sentences stand for facts F_1, F_2, and F_3, respectively.

Consider the following seven sentences.

1. Fa. By assumption.
2. a ≠ b. By assumption.
3. Gb. By assumption.
4. a = the unique thing x such that (x = a and Fx). Follows from 1, by G_1.
5. a = the unique thing x such that (x = a and x ≠ b). Follows from 2, by G_2.
6. b = the unique thing x such that (x = b and Gx). Follows from 3, by G_1.
7. b = the unique thing x such that (x = b and x ≠ a). Follows from 2, by G_1.

1 and 4 stand for the same fact, namely F1. 2, 5, and 7 all stand for the same fact, namely F_2. And 3 and 6 stand for the same fact, namely F_3. "The unique thing x such that (x = a and Fx)" co-refers with "the unique thing x such that (x = a and x ≠ b)", since both expression refer to a. Therefore, 5 is what results when the definite description in 4 is replaced with a co-referring

definite description. Therefore, 4 and 5 refer to the same fact. Since they refer, respectively, to F_1 and F_2, it follows that F_1 and F_2 are the same fact. "The unique thing x such that (x = b and x ≠ a)" co-refers with "the unique thing x such that (x = b and Gx)," since both expressions refer to b. Therefore, 7 is what results when the definite description in 6 is replaced with a co-referring definite description. Therefore, 6 and 7 refer to the same fact. Since they refer, respectively, to F_3 and F_1, it follows that F_3 and F_1 are the same fact. Now, if G_4 is wrong, then definite descriptions are not referring terms, which case 5 is not what results when the definite descriptions in 4 is replaced with a co-referring expression, and 7 is what not results when the definite description in 6 is replaced with a co-referring expression. But assuming that G4 is right, we've established that $F_1 = F_2$. Therefore $F_1 = F_2 = F_3$, and "Fa" therefore co-refers with "Gb" and also with "a ≠ b." So, assuming that definite descriptions are what they appear to be, viz. referring terms, it follows there is only one fact.

As Neale points out, this result, if it is in fact true, is astonishing. But, as Neale then points out, G_2 allows us to go further. Let S and S* be *any* two distinct statements that are true. (So S and S* needn't have subject-predicate form.) According to G_2, S and be put in subject-predicate form, and so can S*. Given any existing individual a, S can be rendered as "a is an x such that S"; and given any existing individual b, S* can be rendered as "b is an x such that S*." And once that is granted, it follows, by the foregoing argument, that S and S* stand for the same fact. It follows that, given G_1-G_4, any two true sentences stand for the same fact.

Evaluating Neale's argument: The problems with this argument are very similar to the problems with Davidson's. Consider line 4 of Neale's argument:

4. a = the unique thing x such that (x = a and Fx).

If this sentence is read ***attributively*** (in other words, if the definite descriptions in them are treated as quantifiers, as opposed to referring terms), then it is logically equivalent with

1. Fa,

in which case those two sentences can therefore be seen as standing for the same fact. But if 4 is read ***referentially*** (in other words, if the definite descriptions in them are treated as referring terms, as opposed to quantifiers), then 4 stands for the same fact as:

4_R: a = a,

in which case 1 and 4 cannot plausibly be seen as standing for the same fact. For exactly similar reasons, 5, 6, and 7 must be read attributively if they can plausibly be regarded as standing for the same facts as, respectively, 2, 3, and 4. In general, G_1 is false if 4–7 are read referentially, 4–7 must therefore be read attributively. Bottom line: Neale's argument fails if 4–7 are read referentially. It also fails, for obvious reasons, if they're read attributively.

Sorites paradox*:* Any paradox like the following. Somebody with zero dollars is poor. Giving one dollar to a poor person won't make him stop being poor. So, no matter how many times you give a dollar to a poor person, he won't stop being poor. Therefore, a person who has been given $500,000,000 dollars is poor. A sorites paradox always consists of a base-clause (e.g., "a person with zero dollars is poor"), an inductive clause (e.g., "if a person with n dollars is poor, so is a person with n + 1 dollars), and an obviously false conclusion that seems to follow (e.g., "a person with a billion dollars is poor")."

Commentary: My solution to this paradox: Statements of the form "x is rich" are elliptical for statements of the form "x is rich compared to y," where y is either some person or some benchmark. When the benchmark is identified, the underlined sentence is seen to be false. If Frank is the person/benchmark in question, and Frank has $1,000,000, then there is a number n (namely, 999,999.7) such that a person with n dollars is not rich and such a person with n dollars is rich.

Special compositionality*:* A special form of compositionality, which we will refer to as "special compositionality," concerns reference. According to special compositionality, what a referring term refers to is a function of what its parts refer to. "The author of *War and Peace*" refers to one person. "The author of *Crime and Punishment*" refers to someone else. Why do those

expressions refer to different people? Because *War and Peace* doesn't refer to the same book as *Crime and Punishment.* This shows that what "The author of *War and Peace*" refers to depends on, and is thus "a function of," *War and Peace* refers to, and it shows the same thing *mutatis mutandis* to be true of "The author of *Crime and Punishment.*" Given obvious extensions of this argument, it follows that *in general* what referring terms refer to depends on what their parts refer to. See "compositionality."

Speech act[512]: Assertions are speech acts, and so are promises, orders, exhortations, and so on. A rough definition of "speech act" is "any case of somebody's deliberately uttering or otherwise ***tokening*** a sentence."

A person may produce a burst of noise, and *know* it to be a sentence, but not thereby have performed a speech act. Knowing that you find loud "s"-sounds completely intolerable, I shout out "snow sometimes symbolizes sorrow" while in your presence. Even though I knew the noises I was making to be a sentence-token, it wasn't my intention in producing them to token a sentence, and I therefore wasn't performing a speech act.

Surd*:* A "surd" is anything that cannot coherently be supposed to exist. Thus, square circles are surds, and so are events that pre-date themselves.

Though pedagogically effective, this definition is logically defective. There exists nothing that cannot possibly exist. (There exists no object x such that x cannot possibly exist. Thus, there is nothing to which, in saying that "surds" are things that cannot possibly exist, we are saying that "surd" refers. (There exists no x such that, in saying that the word "surd" refers to x, we are saying that "surd" refers.)

The word "surd" can be non-defectively defined only in a circuitous manner. For any property phi, the sentence ‹anything having phi is a surd,› means that the laws of logic prohibit phi from being instantiated. (Given that not all sentences containing the word "surd" have the form ‹anything having phi is a surd,› this definition needs fine-tuning. But whereas the problems with our initial definition were substantive, the problems with this definition are merely technical.)

Supervenience*:* Two organisms that are microphysically exactly alike are also biologically exactly alike. Thus, the microphysical locks the biological into place. Put another way, the biological supervenes on the microphysical.

Two statues that are physically exactly alike are *ipso facto* aesthetically exactly alike. Thus, the aesthetic *supervenes* on the physical. Two situations that are physically and psychologically are *ipso facol* morally exactly alike. Thus, the moral supervenes on the physical and the psychological (taken together).

Synthetic knowledge*:* Knowledge is synthetic when what one knows is given by a synthetic truth. See "synthetic truth."

Synthetic truth*:* A statement is synthetic if it isn't analytic. See "analytic" and "synthetic knowledge."

Tautology/truism*:* A tautology (or truism) is a trivially true statement—for example, "there are three feet in a yard" or "Tuesdays come after Mondays" or "there are two weeks in a fortnight."

Theory of Descriptions*:* The position, first advocated in 1905 by Bertrand Russell, that definite descriptions are not devices of reference and are instead quantifiers. "The inventor of bifocals" seems to refer to Benjamin Franklin. But "the inventor of bifocals snored" is true iff exactly one thing invented bifocals, and any such snored. "Smith kicked the inventor of bifocals is true iff exactly one thing invented bifocals, and Smith kicked any such thing. What this shows is that, in virtue of containing an occurrence of "the inventor of bifocals," an expression encodes an existence-claim. This, in its turn, suggests that "the inventor of bifocals," and by parity of reasoning all other definite descriptions, are quantifiers, not singular terms.

Truth*:* A true proposition is said to be one that "corresponds to the facts." But this statement is useless, since the expression "corresponds to the facts" is as obscure as the expression "true."

Given any proposition, there is some condition that is fulfilled if, and only if, that proposition is true. Conditions may be identified with properties. Therefore, truth may be identified with instantiatedness, and propositions (i.e., the things that have the property of being true) are properties (or, strictly speaking, sets there). See Chapter 3.

Many philosophers (e.g., Ayer, Hempel, Strawson, Wittgenstein) were, at least at certain junctures in their careers, of the view that "truth" (and "true"

and other such terms) were meaningless. There is, it was said, no difference between saying “the sentence ‘snow is white’ is true” and “snow is white”; therefore, the words “it is true that” (and “it is a truth that,” and so on) are useless. Wittgenstein was vehement in his advocacy of this viewpoint, which is known as “deflationary” (or the “redundancy”) analysis of truth. The deflationary analysis is false. If “white” meant “black,” but the English language were otherwise unchanged, “snow is white” would be true in a world where snow was black and false in a world where snow was white. So

(1) “snow is white”

is not equivalent with

(2) “the sentence ‘snow is white’ is true.”

(1) makes a statement about snow; it doesn’t say anything about language. (2) makes a statement about language; it doesn’t say anything about snow. (2) concerns the *word* “snow.” But snow isn’t a word; it’s a substance.

This brings us to next problem with the deflationary analysis. (2) is a true statement. There is no way to make that statement without using the word “true” or some equivalent. If you just say “snow is white,” you are not, as we just say, saying anything about the word “snow.” (What you are saying, in saying “snow is white,” would be true in a world where there were no words, but where snow was white.)

Also, any sentence containing a quantifier contains an occurrence, if only a phonetically unrealized one, of the word “true” or of some equivalent. To say

(SD) “some dolphins can read”

is to say that

(SD) for some object x, the proposition *x can read* is true.

Truth-value*:* A true sentence has the truth-value *true*, and a false sentence has the truth-value *false*. The truth-values *true* and *false* can be thought of as, respectively, the property of being true and the property of being false.

If two sentences are both true or they are both false, they have the same "truth-value." So "1 + 1 = 2" and "2 + 2 = 4" have the same truth-value—they have truth-value *true*. And so "1 + 1 = 3" and "2 + 2 = 5" have the same truth-value—the truth-value *false*.

Universal generalization*:* A sentence expresses a "universal generalization" if the ***main connective*** of that sentence is a ***universal quantifier*.**[513] A connective is an expression that, when given a sentence, or open sentence, or ordered n-tuple of (open) sentences, yields a new sentence. Thus "and" is a connective because, when given the pair of sentences ("grass is green," "snow is white"), it yields the sentence "grass is green and snow is white." (In what follows, "sentence" will be short for "sentence or open sentence of ordered n-tuple of sentences or open-sentences.")

Commentary: Many statements that *seem* to be conditionals are not, and are instead universal generalizations. For example *if a thing likes apples, it is friendly*, is a universal generalization, not a conditional. The reason: it's identical with the proposition: *all things are such that they are friendly if they like apples.* Put another way, it's identical with: *for any x, if x likes apples, x is friendly*. Thus, the "main connective" of the sentence "if a thing likes apples, it is friendly" isn't "if," and is instead "for any object x"; and for this reason, it's a universal generalization, not a conditional.

Utilitarianism*:* The doctrine that an act is morally correct to the extent that it promotes human welfare, and morally incorrect to the extent that it undermines it. See "consequentialism."

Vagueness*:* See "ambiguity vs. vagueness."

Valid*:* An argument is valid iff the conclusion cannot possibly be false, if premises are true. A valid argument can have false premises. (For example: Premise: All people have feathers. Premise: All things that have feathers drink to excess. Conclusion: All people drink to excess.) A valid argument with true premises is a *sound* argument. (For example: Premise: JMK has a car. Premise: All cars are motorized vehicles. Conclusion: JMK has a motorized vehicle.)

Validate*:* See “interpretation.”

Virtue theory*:* The doctrine composed of the following two assertions: (i) a morally correct life is one in which one flourishes. (ii) One’s own flourishing depends constitutively on one’s having character traits (or “virtues”) that conduce to the flourishing of others.[514]

[1] Expressions that are italicized and boldfaced are very important ones whose meanings must be known. In each case they are defined twice: once in the main text and once in the Appendix.

The term "sensory observation" refers to any instance of sight, hearing, touch, etc.

[2] Fodor makes this point, in an exceptionally clear and forceful manner, in his (1981) essay “The present status of the innateness controversy,” which is one of the two or three best philosophical papers I have ever read. What I say in Chapter 9 of the present book is heavily informed by my reading of that paper.

[3] “BH” stands for “base times height.” It is my practice to give names to sentences that are frequently referred to, since it eliminates unnecessary verbiage from my prose and (partly, though not primarily, for that reason) makes it easier to follow.

[4] All principles express dependence relations, though this isn’t always obvious. Consider the principle that: (MD) military dictatorships tend to be less responsive than democratically elected governments to the needs of those whose welfare they are duty-bound to promote.

MD is to the effect that, human psychology being what it is, *x*'s being a military dictatorship depends on *x*'s being less likely than a democratically elected government, other things being equal, to meet the demands of those whose interests it is duty-bound to promote.

Because all principles express dependence relations, Reason #1 and Reason #2 ultimately coalesce. A principle is not merely a general truth. That's why "any given tiger in North America is in some zoo" is not a principle. A principle must be intrinsically, as opposed to contextually, *explanatory*.

To be sure, "any given tiger in North America is in some zoo" *can* be explanatory---but only under rather special conditions. The way in which "a body's mass increases as it accelerates" is explanatory is clearly not so context-dependent---or, even if it is context-dependent to the same degree, it isn't context-dependent in the same way.

Unfortunately, since this manuscript is already overdue, these vague remarks are the best I can do in the way of making it clear how a truth must be explanatory for it to be a principle as opposed to a mere general truth.

[5] Why "probably not always"? Because, it seems to me, there are some non-empirical facts (e.g., nothing can be located, in its entirety, in two disjoint places at a given time) of which one can have knowledge *without* having to have analyzed anything. Also, it seems to me, that *in order* to analyze a concept and thus to obtain knowledge through conceptual analysis, one must *already* have non-empirical knowledge.

Some individuals hold that there is non-empirical knowledge that does *not* concern the structures or interrelations of concepts and therefore is *not* analytic, as we are using this term. For example, Kant (1787) held that there is non-empirical, non-analytic knowledge. In Chapter 10, it is argued that all non-empirical knowledge is indeed analytic.

Actually, this last statement isn't really accurate, for reasons I discuss in my (2006) article, *The analogue-digital distinction and the cogency of Kant's transcendental arguments*.

[6] And, the sociology of this discipline being what it is, many "leaders in the field" compare unfavorably in respect of their philosophical prowess than people of only slightly above average philosophical acumen. People of talent are seen as a threat and are weeded out, seldom making their way into graduate school, let alone managing to become professors.

[7] Actually, to echo what I said a moment ago, everything that I'm saying about "neophytes" is true of senior professors. It's a rare philosophy professor who has anything to say. 99 philosophers out of a 100 make their living by tearing down the works of other philosophers. Any obscurity is seen as an opportunity to straw-man the author of that text.

[8] The "‹" and the "›" are known as ***quasi-quotation*** marks. this context they may be treated as quotation marks. But, strictly speaking, they are not quotation-marks, and the differences between them and quotation marks are important in some contexts. An example of a variable is the expression "x" in "for any object x, if x has three cars, then x has more than one car." An example of an anaphoric pronoun is the "he" in "if a man owns more than three cars, he owns at least one car." It's obvious that anaphoric pronouns are similar to variables, if not identical with them.) What is the purpose of quasi-quotation marks? Consider the expression: "if *P* and *Q* and *R*, then *R*." That expression isn't true or false. So the sentence "For any propositions *P*, *Q*, and *R*, the sentence 'if *P* and *Q* and *R*, then *R*' is true" is *not* true. But the sentence "for any propositions *P*, *Q*, and *R*, a true sentence results when, given the open-sentence 'if *P* and *Q* and *R*, then *R*,' the variables are uniformly replaced with arbitrary sentences" is true. And (A) is a concise way of saying this. In general, quasi-quotation marks make it possible to express generalizations about sentences in a concise way. Quine was the first to see the need for quasi-quotation marks and duly introduced them in his (1941) book *Mathematical Logic*.

Why, incidentally, did I choose to explain the meanings of the terms "variable" and "anaphoric pronoun", not by defining them, but by giving examples of sentences in which they are used? Three reasons. First, to my knowledge, no viable definitions have been given, at least not in print. (In yet to be published work, I put forth what I believe to be definitions of them that are not totally implausible.) Second, any accurate definition of these terms would be extraordinarily hard to follow for anyone who didn't have a lot of background in semantics. accurate definitions are very hard to produce. This is a consequence of the fact that, with some trivial exceptions, *all* definitions are extremely intricate.

The exceptions include ones where, given an expression belonging to one language, one is merely identifying its counterpart in some other language. They also include stipulative definitions, e.g. "let 'blerk' mean *triangle*." And there do seem to be *some* cases where, non-translational, non-stipulative definitions are not to so hard to produce, e.g. "a 'triangle' is a three-sided, straight-edged, closed, planar figure."

But a few comments about that definition are in order:

(i) That definition embodied an *analysis* of the word "triangle" and, for that reason, wasn't easy to produce. (I didn't produce it; I merely repeated it.)

(ii) Nobody learns what "triangle" means by being given such a definition.

(iii) A definition (or analysis, or whatever it is) is useless

unless knows what is meant by terms (e.g. "planar", "closed") whose meanings one is unlikely to know unless one knows the meaning of "triangle."

(iv) That definition fails to fix the meaning of the word "triangle" and is thus not a viable definition. This is because the terms "straight", "closed", and so on, do not themselves have unambiguous meanings outside of artificial contexts. They are actually variables, not constants, whose meanings are specific, either implicitly or explicitly, by the dialectical context. The reasons for this are given in Chapter 1 of my book *Analytic Philosophy*, in the section titled "mathematical logic," which is the at the very end of that chapter.

[9] . In Chapter 3 of *Analytic Philosophy*, I provide a detailed answer to the question "what properties must an entity have if it is to be capable of being whatwe are thereby affirming in uttering declarative sentences?" In Chapter 3 of the present work, I put forth a truncated version of my answer to that question.

[10] See Kment (2007).

[11] After attending a lecture that Kment gave in 2006, I asked him in private why (3) isn't the right analysis of (1). His response was that my analysis was self-evidently false and that my question didn't deserve a response. (Out of the 20 or so people who publicly asked Kment questions about his lecture, only one audience-member, Bert Bandman asked a question that didn't assume the truth of Lewis' view. Bandman asked whether there

were any alternatives to the view that statements such as (1) demanded that we posit the existence of alternative universes. Kment rolled his eyes, looked at his watch, and then said with a sigh: “yes, but none that are worth considering.” Audience-members snorted at Bandman and laughed heartily at Kment’s witty and informative reaction to his question. They then proceeded to ask questions which assumed the correctness of Kment’s view and were sometimes prefaced by admiration for the patience and decency Kment had shown in his response to Bandman‘s deeply inappropriate question.

Apparently, only one audience member, besides myself, had any reservations about granting the existence of concrete (as opposed to purely fictitious) alternative universes. Which I found odd, given that we were at a philosophy conference, not a Grateful Dead concert.

[12] See Lewis (1973, 1984).

[13] Interestingly, British actors have a similar view. “To think like Hamlet“, they say (this is not a quotation), “you must first *act* like Hamlet. Fake it till you make it. If you fake it long enough, you won’t be faking it anymore: the mask you put on will now be your real face.“ American actors have the opposite view: “In order to *act* like Hamlet”, they believe (again, this is not a quotation), “you must first *be* Hamlet: you must have his thoughts, feelings, and so on.” Actors who follow this principle are known as “method actors.”

[14] The psychopath must not be confused with the cynic. A cynic is somebody who himself has strong values but believes that, for the most part, others either don't have them or do have them but fail to comply with them. (Schopenhauer was a cynic; so was La Rochefoucauld. These thinkers are heroes of mine.) The cynic is, in many ways, the antithesis of the psychopath.

The word "cynic", I must point out, is ambiguous. Sometimes the word "cynic" refers to somebody who doesn't have values or has serious reservations about whether it's worthwhile to comply with them. But I will

use the word "cynic" to refer to somebody who is skeptical, not about whether values are worth having, but about whether people are trying to comply with them.

[15] See Malcolm (1956).

[16] See Ryle (1949).

[17] See Russell (1919, p. 168) for a similar argument. The argument that I just put forth was, in respect of its substance and its wording, deliberately modeled on Russell's argument.

[18] This argument is an adaptation of one put forth by Peter Geach (1972). Geach was attempting to refute a doctrine called emotivism. But the structure of Geach's argument is the same as the structure of mine.

[19] See Austin (1962) for Austin's own rendition of his argument.

[20] "AA" is short for "Austin's argument."

[21] At least in English, unless English were changed in profound ways. It's hard to believe that, in other languages, embedding a sentence *P* in a conditional of which it is the antecedent changes *P*'s meaning. So, in the aforementioned respect, English is presumably representative of all human languages, a consequence being that indirect arguments would be inexpressible, in any human language. But, so far as we know, they are expressible in all of them, without any of them having to undergo any changes at all.

[22] *"Mutatis mutandis"* means: *provided that the needed changes are made*. Consider the statement:

(i) "If I make sure that I exercise and get plenty of rest, and everybody else does the same thing, the world will be a better place."

(i) is ambiguous. It could mean:

(ii) if I do my best to make sure that I exercise and get plenty of rest, and everybody makes sure that I exercise and get plenty of rest, the world will be a better place.

Or it could mean:

(iii) If I do my best to make sure that I exercise and get plenty of rest, and each other person makes sure that he or she exercises and gets plenty of rest, the world will be a better place.

Unlike (i), the following sentence *unambiguously* has (iii) for its meaning:

(iv) if I make sure that I exercise and get plenty of rest, and everybody else does the same thing *mutatis mutandis*, the world will be a better place.

Sentences that are ambiguous in much the same way as (i) often occur in analytic philosophy, and it's often unclear how they are to be disambiguated. But the expression "*mutatis mutandis*" makes it easy to disambiguate them.

[23] Creating the right mechanisms may involve an enormous amount of theoretical knowledge—a N.A.S.A. engineer may know a lot more about physical law than many a professor of theoretical physics. Whether one is an engineer as opposed to a physicist is a function not of how much one knows about physics, but about one's attitude towards that knowledge. The engineer is interested in physical law only to the extent that it helps build bridges and tanks. The physicist is interested in bridges and tanks only to the extent that it helps identify physical laws. That is why even though Newton was one of the greatest physicists of all time many a contemporary engineer knows more about physical law than he did.

[24] He says this in his (1978) book *Truth and Other Enigmas*, in the chapter titled "Can analytic philosophy be systematic, and ought it to be?"

[25] This is now contested. Some philosophers hold that there are coherent statements—statements that don't embody internal inconsistencies—that cannot possibly be true. See Putnam (1975) and Kripke (1980). In Chapter 8, it is argued that this is false.

[26] "FL" is short for "Frege's legacy."

[27] See Frege (1980). Russell (1905, 1914, 1917, 1918, 1927, 1948) did a very good job of clarifying and developing this viewpoint.

[28] Russell (1903).

[29] The argument about to be given was put forth by Russell (1920).

[30] “TM” is short for “three minutes.”

[31] “UP” is short for “uninstantiated property.”

[32] See Pap (1949). Also see Carnap (1932) and Schlick (1934).

[33] “TP” is short for” “tall person.”

[34] Frege used the word “concept,” not” “property.”

[35] Russell makes this point in his (1948) *History of Western Philosophy* and in many other works of his.

[36] Here we'll add to what was said about this principle in the introduction.

[37] “LP” is short for “Larry perception.”

[38] “LS” is short for “Larry sentence.”

[39] An expression is “semantically simple” if it doesn’t consist of other expressions. So “the person in the cockpit at the time of the call” is semantically complex, but “Socrates” and “Larry” are not.

[40] See Russell (1917, Chapter 10), Barwise (1983), Kuczynski (2007).

[41] See Brentano's *Psychology from an Empirical Standpoint.*

[42] This apt way of summing up special relativity is due to Howard Gardner (1976).

[43] Empirical discoveries in psychology and physics do play a part in the dissolution of philosophical problems. But usually when some empirical discovery sheds light on a philosophical issue, it is by prompting philosophers to produce the relevant analytic knowledge. Certainly a lot of good philosophy was *prompted* by what Einstein and Freud had to say, but the ensuing philosophical insights were ones that, theoretically, could have been arrived at independently.

[44] And Frege's knowledge of the meanings of specific German sentences, while obviously empirical, wasn't materially implicated in the points he was making. He could have made those points had he not known a word of German and had been writing in, for example, Hindi or Spanish.

[45] Wittgenstein (1958).

[46] Wittgenstein (1922).

[47] Epistemology is the theory of knowledge.

[48] See Fodor's (1990) book *A Theory of Content* and his (1998) book *Concepts.*

[49] See my (2007) book *Conceptual Atomism and the Computational Theory of Mind.*

[50] Although this position is not as popular as Fodor's, there are philosophers besides myself, e.g. Laurence Bonjour (1998) who also advocate it.

[51] Incidentally, the very fact that JK is to be adjudicated in a non-observational manner shows that JK is false. Supposing JK correct, the fact it describes is that one statement is inconsistent with another; and that fact is obviously causally inert, that being why it cannot be observed.

[52] In this context, the word "sentence" will stand for both ***sentence-types*** and ***sentence-tokens;*** i.e., for sentences *per se* and also for people's utterances and inscriptions of them.

[53] See Frege (1918).

[54] In the sentence, "*x* is a certain way," the word "is" doesn't denote the relation of identity. The "is" is being used ***predicatively.*** It's being used in the way it is in the sentence "Smith is happy," which obviously doesn't say that Smith is *identical* with happiness; and not in the way it's being used in the sentence "Smith is the guy who stole my car." See Chapter 16, Section 1.1 for further discussion.

[55] Actually, we'll see in Chapter 4 that propositions are *sets* of properties. But this is a nuance. And it's an irrelevant nuance, since, as we'll see in Chapter 4, sets *are* properties and therefore propositions are identical with sets. This must be understood aright. The property of being a bird is *not* identical with the set of all birds—contrary to what Quine (1960, 1970) and other anti-Platonist philosophers have held.

[56] And so far as it *is* up to us how it is, it's up to us in a *causal* sense. It is in virtue of what people do, not what opinions they hold, that certain states of affairs hold as opposed to others. Even though, on the basis of an opinion, one may perform acts that wrench the world into a certain form, one's opinion *per se* is innocuous: The world is the way it is regardless

of that opinion.

Also, it's not a matter of opinion what opinions we have; it's a matter of fact. We have the opinions that we have and, so far as we know ourselves, our opinions as to what those opinions are track those pre-existing facts. Our second-order opinions—i.e., our opinions about opinions—no more make our first opinions be the way they are than our first-order opinions about the stars make the stars be the way they are. Our second-order opinions may induce us to engage in acts that result in the modification of our first opinions, whereas our opinions about the stars cannot possibly induce us to engage in acts that change the stars. But that's irrelevant, since it doesn't faze fact that it is never *in virtue of* what one's opinions are that the world is, or is not, in any respect, a given way.

[57] This is the branch of philosophy that aspires to clarify the concepts that are used in formal logic but that themselves resist formal treatment; e.g., ***entailment, formal truth, informal analytic truth, possibility, logical form,*** etc. See Section 7.0 for further discussion. Most of what is said in Part 1 of the present book falls within the jurisdiction of philosophical logic.

[58] "Acoustical" means "noise-related," and "morphological" means "shape-related."

[59] As a matter of exegetical fact, I have in no way exaggerated Wittgenstein's claims.

[60] He hedges by "roughly speaking, meaning is use." He never explains the hedge. I've omitted the first two words, since they serve no identifiable purpose.

[61] "WA" is short for "Wittgenstein's argument." This is not a quotation. There is no one passage where Wittgenstein argues on behalf of his thesis that meaning is use. His arguments are dispersed over many pages,

in most cases being presented in the form of aphorisms and rhetorical questions. Quine (1960) advocates the position about to be presented, and he states it very crisply. Some of the locutions that occur in WA are found in Quine (*op. cit.*), and some, though not (so far as I know) used by Quine himself, are Quinean.

[62] The argument being put forth is found in precisely this form in Quine (1960). It is also Wittgenstein's position. But Wittgenstein doesn't state it as crisply as Quine.

[63] A person who *did* know that language would agree that I hadn't said anything *after* that person had learned that I didn't know what they meant. That's why if those noises were offensive, that person would cease to be offended *after* he learned that I didn't know what they meant.

[64] There are different views as to what it is. But, as is universally conceded, they are not themselves sentences. See Chapter 4.

[65] In Chapter 4, each of these terms is defined and the concept expressed by it is explained and critiqued.

[66] In Chapter 4, each of these terms is defined and the concept expressed by it is explained and critiqued.

[67] "TA" is short for "Tractarian argument." TA is a paraphrase, not a quotation. But some parts (e.g., BG) are direct quotations.

[68] *Tractatus Logico-Philosophicus,* Section 4.003. I am using the Pears/McGuinness translation.

[69] *Tractatus* Logico-*Philosophicus,* Section 4.003. I am using the Pears/McGuinness translation.

[70] Ryle (1949).

[71] Ryle (1949).

[72] A conjunct is one of the sentences composing a conjunction. A conjunction is a sentence that consists of two or more sentences that are joined by an “and” (or some other comparable expression; e.g., “but”).

[73] “CT” is short for “central thesis.”

[74] “Corollary” is a synonym of “consequence.” A corollary of nation *X*’s having all its troops in Nation *Y* is that it can’t station any troops in Nation *Z*.

[75] See Wittgenstein (1922), Carnap (1932, 1934), Schlick (1932, 1934), Hahn (1933), Stevenson (1937), Hempel (1950), Ayer (1952).

[76] “Iff” means “if and only if.” “S_1 iff S_2” means “if S_1 is true, it follows that S_2 is true; and if S_2 is true, it follows that S_1 is true.” Thus, “S_1 iff S_2” means that the truth of S_1 is both ***necessary*** and ***sufficient*** for the truth of S_2.

[77] In addition to being false, the statement: (VE) “all meaningful non-tautologies are capable of being observationally verified” isn’t a tautology, since it’s debatable (and false, as we saw). Given that VE isn’t a tautology, one who accepts VE must assume it verifiable if not meaningless, and must therefore assume it verifiable. But VE is unverifiable for the same reason as the statement “all metal expands when heated” would be. Any attempt to show otherwise would be self-refuting. One can’t try to show that VE is verifiable without first assuming it meaningful. But in assuming this, one is assuming that unverifiable non-tautologies *can* be meaningful and, therefore, that VE is false.

[78] The term “falsificationism” is associated with Karl Popper (1902–1994). But, contrary to what is sometimes alleged, Popper didn’t advocate the doctrine to which that expression refers in this context. He did not hold that a statement had to be falsifiable to be *meaningful*. Popper accepted the doctrine that for a statement to be *scientific* is for it to be falsifiable. (Both doctrines are known as “falsificationism.” Hence the mix-up.) He makes this clear in his (1983) book *Realism and the Aim of Science*.

[79] Carl Hempel (1905–1997) made points similar to the ones just stated. See Hempel (1965) Chapter 3. In Section 7.1, we’ll consider another, subtler argument of Hempel’s for the position that not all knowledge is observation-based.

[80] In Section 4.8 we’ll question this assumption.

[81] Technically, ink marks, noises, etc. are sentence tokens, not sentences. But we can set that aside here.

[82] Carnap (1934).

[83] He didn’t put it this starkly. And he probably wouldn’t have accepted it—at least not for long—if it had been presented to him this starkly. But that doesn’t mean that it wasn’t his view. It was.

[84] This is the view of Wittgenstein (1958), W.V.O. Quine (1960), Hartry Field (1977), Robert Brandom (1994), and John McDowell (1998).

[85] “NT” is short for “Newton’s theory.”

[86] The “‹” and the “›” are known as ***quasi-quotation*** marks. See the Preface, Footnote 8, for an explanation of what they are. In this context, they may be treated as quotation marks.

[87] Strictly speaking, it doesn't quite say this. S would say how bodies x_1 and x_2, having masses m_1 and m_2, would behave *to the extent that* their behaviors would be functions of their having those masses and being that distance from one another. But other factors might obviously be involved; e.g., x_2 might be near some third body x_3 whose gravitational pull was much stronger than x_1s' or it might be under the influence of some non-gravitational force.

[88] See Hempel (1965, Chapter 3), Sheffler (1981, Part II).

[89] Wittgenstein, himself LP's founding father, *did* try to make a case that LP was fundamentally misguided, and not just technically faulty. He spent 20 years trying to make a case for this, and his efforts were embodied in a now famous book known as the *Logical Investigations.* But the doctrines put forth in the *Investigations* are quite as incoherent as LP itself. See Chapter 5.

[90] I'm quoting from the McGuiness/Pears translation.

[91] Russell (1941, pp. 14–15).

[92] In an essay on Russell's work, published in a volume titled *The Philosophy of Bertrand Russell* (edited by Paul Arthur Schilp), W.T. Stace (1886-1967) made this point.

[93] In her very good book *Physics and the Philosophers,* L. Susan Stebbing makes similar point. Stebbing is discussing an argument given by Arthur Eddington (in his book *The Nature of the Physical World*) that is similar to Russell's.

[94] "GA" stands for "Goodman's argument." This is not a quotation.

[95] What follows wasn't said by Goodman; I'm developing his point.

[96] See Mates (1952) for some arguments that, in my judgment, embody a failure to take these points into account.

[97] Martin Heidegger did in fact say this, but the fact that one person said it doesn't mean that the logical positivists were not attacking a straw man.

[98] Wittgenstein says that the alphabet is a picture "in the ordinary sense"of the corresponding sounds—that, for example, "*k*" is a picture, "in the ordinary sense," of the corresponding sound. That isn't true. The relationship between "*k*" and the corresponding sound is conventional, not iconic.

[99] Frege, who held the same view, said that "propositions are true facts." Frege's view isn't quite accurate. False propositions aren't facts. It isn't a fact *that the moon is made of cheese.* Bearing in mind that false propositions aren't facts, let's suppose, with Frege, that true propositions *are* facts. Facts are, in at least some cases, spatiotemporal entities. The fact of my typing right now consists of various displacements of mass-energy. So if true propositions were facts, then some propositions would be spatiotemporal entities, whereas false ones would categorically fail to be such entities. But it would be theoretical arbitrariness of the worst kind to hold that some members of the class of propositions—which is obviously a unified, highly integrated class—were spatiotemporal, whereas other members of that same class were not.

[100] Goodman (1976) discusses the concept of a digital representation as well as the concept of an analogue representation. We will discuss that concept in Chapter 5.

[101] The differences between iconic and linguistic representation are

discussed at length in Chapters 21–23 of my (2007) book *Conceptual Atomism and the Computational Theory of Mind: A Defense of Content-Internalism and Semantic Externalism.*

[102] "Grass is green and snow is white" is the result of ***conjoining*** "snow is white" and "grass is green," and "Grass is green or snow is white" is the result of ***disjoining*** them.

[103] A statement is vacuously true if it's trivially incapable of being counter-exampled. An example is: "every human being who lives on Neptune plays the accordion" In order for that sentence to be false, there would have to be a human being on Neptune who didn't play the accordion. Since there are no human beings on Neptune to begin with, there are none who don't play the accordion.

[104] Berkeley made this point very clearly in his (1734) work *The Principles of Human Knowledge.*

[105] This is an adaptation of the Pears/McGuinness translation. I've replaced each occurrence therein of "proposition" with "sentence," for the reasons given earlier. Apart from that, I've made no substantive changes; but I've touched up a few of the sentences, in ways that do not modify their meanings, so as to eliminate possible sources of confusion that my editor pointed out to me. The italics are Wittgenstein's own.

[106] I suspect that it was his awareness of this fact that led him to end his book by saying that everything in it is nonsense.

[107] Wittgenstein's (personal and philosophical) mentor, Bertrand Russell, was at that point the world's authority on paradoxes of this kind. Russell's work on these puzzles is by no means obsolete. Wittgenstein was in daily contact with Russell during much of the five year period preceding his

writing the TLP.

[108] A historical point: In his (1734) book *Principles of Human Knowledge*, George Berkeley made virtually every point that Wittgenstein made in the *Tractatus.* Moreover, in the *Principles* and in other works (e.g., *An Essay Towards a New Theory of Vision*) Berkeley made a number of cogent, insightful, and strikingly modern points, all of which are absent from the TLP. Berkeley's important work will be discussed in Chapter 13.The one claim that is clearly present in the TLP that is clearly absent from the *Principles* is the thesis that for a sentence to be meaningless is to be syntactically ill-formed. Berkeley claimed that for a sentence to be meaningless is for it to be unverifiable; i.e., incapable of being definitively shown to be true on the basis only of what one's senses tell one. In other words, Berkeley was a verificationist.

Of course, the TLP advocates verificationism. But verificationism is inconsistent with one of the other main claims of the TLP, namely, that meaninglessness is categorically the result of bad syntax. I leave it to the reader as an exercise to demonstrate this.

[109] Such references are found in the work of Russell (1918), Wittgenstein (1922), and Hempel (1950).

[110] A "logically perfect language" isn't the same thing as a formal language. But, like many others in his time and ours, Hempel didn't see this. The thesis that meaningful sentences are those that can be expressed in logically perfect languages is, I am certain, an expression of this confusion.

[111] Hempel (1952, 1965).

[112] "HA" stands for "Hempel's argument."

[113] Hempel (1952) does an extremely good job of making it clear what conditions an object must fulfill if it is to

serve as a metric standard.

[114] To say that Ö2 is an irrational number is to say that there are no whole numbers *p* and *q* such that *p*/*q* = Ö2. This entails that, given a length L_1 of Ö2 units and also a length L_2 of N units, where N is a rational number, there is no rational number M, such that each of L_1 and L_2 can be divided, without remainder, into M units. Two lengths that are related to each other in this way are "incommensurable." It was, it is said, Pythagoras who discovered the existence of incommensurables. Richard Dedekind (1831–1916) and Bertrand Russell (1872–1970) jointly provided the first clear analysis of incommensurability.

[115] As previously stated, a number N is irrational if there are no numbers *p* and *q* such that N = *p*/*q*. See Chapter 7.

[116] Hempel (1952, 1965).

[117] *S* is a set if:

(i) *S* is a collection of objects,

(ii) it is irrelevant to *S*'s identity what relations, if any, hold among those objects,

and

(iii) all that is relevant to *S*'s identity are the identities of those objects.

So a heap of grains of sand is not identical with the set of those grains of sand. For the relations holding among those grains are relevant to the constitution of that heap but not to that of that set. Similarly, the sequence of numbers *1, 2, 3.....99* is not identical with the set containing those numbers. For in that set, those numbers are not ordered, whereas that sequence *is* an ordering of them.

[118] Russell makes this very point in *An Inquiry into Meaning and Truth.*

[119] E.g., Wittgenstein (1922).

[120] The word “nomic” means “having to do with natural law.”

[121] See Armstrong (1993).

[122] Quine (1960) held it. (See also his book *The Roots of Reference*.) In his view, the property of being water is identical with the scattered object consisting of all and only water-molecules. It’s now a popular one. I suspect that this is largely because Quine countenanced it.

[123] Quine (1960, 1981) has this view. So do William of Ockham (see his book *Logical Problems*) and Bertrand Russell (1973).

[124] Plato himself gives this argument in his dialogue *the Parmenides*, and so did Aristotle in the *Metaphysics.* Aristotle gave it its current name.

[125] This is pronounced “one to one.”

[126] See Russell (1920).

[127] See Benacerraf (1965), Dretske (1982), and Fodor (1987, 1990, 1998).

[128] Locke advocates this view. Many who do not explicitly advocate it are committed to it; in other words, the doctrines they do explicitly advocate are false unless conceptualism is true. Every form of empiricism presupposes the truth of conceptualism.

[129] This argument is given by George Berkeley (1710).

[130] Ayer (1952) and Carnap (1934) give this argument, or at least assume its cogency.

[131] The argument about to be given was put forth by Quine (1966).

[132] C.S. Peirce puts forth a similar argument and was the first person, so far as I know, to use the words “type” and “token” in this way.

[133] In any case, anything that is in space-time is an actualized possibility. But not all actualized possibilities are in space-time. The number two, for example, is an instance of the property of being even, and is thus an actualization of the possibility of there being even numbers. But the number two is not in space-time.

[134] M. Oreste Fiocco holds that the property of being round is *distinct* from the property of being a closed, planar, figure of uniform curvature; that the property of being a triangle is *distinct* from the property of being a straight-edged, three-sided figure such that any two of its sides intersect but not all three do; etc. In my judgment, Fiocco's view is not a reasonable one. According to Newton's physics, it is sheer coincidence that an object's gravitational mass coincides with its inertial mass. One of Einstein's great triumphs was to show that gravitational mass *is* inertial mass and that there is therefore no coincidence. Oreste's position is even more arbitrary than Newton's and, whereas Newton had a good excuse for identifying intertial with gravitational mass, Fiocco has no excuse for *not* identifying the property of being round with the property of being a two-dimensional, closed, planar figure of uniform curvature. Relative to the experimental evidence available to physicists before the late 1800's, it was impossible *not* to distinguish between inertial and gravitational mass. Also, Newton's view, though erroneous, was internal to an extremely fruitful scientific theory. Fiocco's is not; it is a wholly unmotivated blunder.

[135] See Grice's classic (1957) article "Meaning."

[136] In this context, the word "people" is meant to refer to sentient creatures of any kind—Martians, dolphins, etc. I wish to leave it open whether non-human creatures have conventions or, therefore, languages.

[137] I say "typically" because not all linguists or philosophers of language consider this division to be a legitimate one. Wittgenstein (1958), for example, thought that semantics collapsed into pragmatics.

[138] The "R" stands for "right," as in "the right analysis."

[139] "CA" stands for "complete analysis."

[140] Given that I'm only indicating in broad terms what the discipline

of semantics is about, I've omitted some steps from this argument. See Kaplan (1989) for a more thorough analysis of sentence-utterances like the one being discussed (viz. sentence-utterances that contain ***demonstratives***—words like "this" and "that.") There's a second reason why my adding the missing steps is unnecessary. The point I'm trying to make (viz. that it isn't always obvious what is literally meant by an expression) is *itself* illustrated by the fact that it may be debated whether U_1 and U_2 have the same meaning.

[141] Kaplan doesn't answer this very question, even though, in some ways, Kaplan does a creditable job of defending this analysis of utterances containing demonstratives. And this question, it seems to me, is much more important than the narrowly semantic question of what the literal meanings of utterances of "that man," "this world," etc., are. Kaplan notes the difficulty (viz. that sentence-tokens that have the very same literal meaning can differ dramatically in what they communicate), and responds by saying (I quote, without omitting relevant contextual material): "Uh oh!" (See Kaplan (1989)), which pretty much exhausts what he has to say about the matter. It also exhausts what Kaplan-disciples, such as Nathan Salmon, have to say about the matter. (See Salmon, 1986, 2005, 2007.)

[142] In 2004, in the journal *History and Philosophy of Logic*.

[143] Note: To understand this section, it is necessary to know what the words "token" and "type" mean when used in semantic discourse. See Chapter 2, Section 3.0 for definitions of these terms. Long story short, a particular occurrence (utterance, inscription, etc.) of an expression is a *token* of that expression, and is thus an *expression-token*, and the thing being tokened is the corresponding *expression-type*.

[144] In this context, the word "object" is being used unusually liberally and thus refers to things (e.g. bursts of noise) that aren't ordinarily characterized as "objects." Basically, "object" is synonymous with "entity" in this context; and, like the latter, the former is, in this context, little more than a cipher.

[145] Lewis (1975).

[146] This is not a quotation; it is a summary of what Griceans say in defense of their position. See Blackburn (1984).

[147] Short for “a Gricean view.”

[148] See Blackburn (1984).

[149] One of the problems with logical positivism was that it assumed, incoherently, that *expressions* could be meaningless. There can be meaningless expression-like entities—but no meaningless expressions. And if, as the logical positivists contended, “there is no free will in a deterministic universe” is truly meaningless, then it isn’t a sentence at all; it is, at best, a sentence-like entity. But it plainly is a sentence. It may be (though I do not myself believe this) that “there is no free will in a deterministic universe” is in some way or other an *incoherent* statement. But an incoherent sentence is one that bears a self-undermining meaning. And there’s a big difference between bearing a self-undermining meaning and bearing *no* meaning. The logical positivists didn’t make this distinction.

[150] Jerry Fodor and Ernie Lepore do a beautiful job of stating this argument and others like it in *The Compositionality Papers.*

[151] See Kaplan (1989) for points very similar to those given in Sections 7.1–7.21.

[152] Mediating concepts are obviously similar to what Frege refers to as ‘senses.’ See Chapter 5. But, not wanting to prejudge the delicate question of whether they’re identical with Fregean senses, I’ve decided to give them a different name.

[153] My personal view is that *ultimately* definite descriptions are

devices of reference. But they are not to be understood in remotely the way that Frege thought. My official policy in this book is that they are to be understood in the way that Russell proposed. Russell's analysis, though not (in my view) a correct answer to the narrowly *semantic* question "are definite descriptions referring terms?" makes it much easier to absorb the larger, more important epistemological, logical, and metaphysical points that it was his real purpose in putting forth the Theory of Descriptions to defend. See Kuczynski (2007), Chapters 1–4 for a defense of the view that definite descriptions are (non-Fregean) referring terms.

[154] There are many reasons for thinking otherwise. I discuss these reasons in Kuczynski (2009), Chapter 6. But my view is that, indeed, they are devices of reference. See Kuczynski (2007), Chapters 2-4.

[155] See Barwise and Perry (1983) and Kuczynski (2006).

[156] "IA" is short for "interior angles."

[157] "BH" is short for "Bill's holdings."

[158] Barwise and Perry (1983) make similar points. See Kuczynski (2007) for a thorough defense of this position. P.F. Strawson (1950) was the first author to put forth a viewpoint of this general kind.

[159] Frege (1954).

[160] Two statements are compatible if neither entails the negation of the other. So "John is in the U.S." is compatible with "John is in Wyoming" but it's ***incompatible*** with "John is not in the U.S."

[161] "WRS" stands for "wrong rule for 'someone.' "

[162] Arguably, not all sentences containing "someone" have the form

‹someone has psi.› For example, “Plato kicked someone” seems not to have that form. For this reason, RFS, at least arguably, isn’t adequate. A more accurate, because more general, definition of “someone” is:

(RFS*) In virtue of having the form ‹...someone...›, a sentence says that the property ...*x*... is instantiated.

Thus, “Plato kicked someone” says that the property being a thing that is kicked by Plato is instantiated. RFS* collapses into RFS. The reason is that any sentence of the form ‹...someone...› can be represented as having the form ‹someone has psi.› For example, “Plato kicked someone” says the same thing as “someone is such that Plato kicked that person.” What we just said about “someone” is true *mutatis mutandis* of all other expressions that are to be defined contextually, which, given the fact that all expressions are to be defined contextually, means that it’s true of all expressions.

[163] “PS” stands for “the property of snoring.”

[164] Frege (1891, 1892b). A research program in linguistics, known as “Montague grammar” is based on developments of these points of Frege’s due to Richard Montague (1930–1971). See Montague (1974).

[165] In this section, we won't discuss molecular propositions (propositions that have other propositions as proper parts; e.g., *snow is white and grass is green*) or quantified generalizations (propositions that, for two classes C_1 and C_2, answer the question "to what extent, if any, do C_1 and C_2 overlap?,”; e.g., "no dogs play poker,” which answers the question "how much does the class of dogs overlap with the class of poker-players?" The non-abbreviated versions of these answers are found in Chapter 3 of my book

Analytic Philosophy.

[166] For expository simplicity, we’re leaving aside niceties relating to tense-markers and place-indicators.

[167] In what follows, we'll set aside irrelevant subtleties relating to the tense-markers occurring in the sentences we're examining, and we'll also set aside subtleties relating to the fact that BF is no longer with us.

[168] This is not a quotation. These are my own words. But they do justice to Frege's position. See Frege (1892).

[169] “FR” is short for “Frege.”

[170] See Chapter 4 for Frege's cogent demonstration that any given sentence must contain at least one expression that doesn't refer to anything. (The point that Frege is establishing has nothing to do with the triviality that there are expressions, e.g. "Elmer Fudd", that don't pick anything out, even though, given they're syntactic properties, they're supposed to do so.

[171] Frege (1892).

[172] “IB” is short for “the inventor of bifocals.”

[173] Actually, according to TD, each of (1) is and (2) is ambiguous between "wide-scope" and "narrow scope" interpretations, (1*) and (2*) being the narrow scope interpretations. (See Russell 1905.) In any case, what's important is that (1*) and (2*) are among the meanings that TD ascribes to (1) and (2), respectively.

[174] Special compositionality is to be distinguished from the more general principle that the meaning (semantics) of a given expression is a function of the meanings (semantics) of its constituents.

[175] More precisely, they don’t co-refer when read *de dicto.*

[176] In any case, supposing that *O* is the prettiest woman in existence, it refers, according to Frege, to *some thing* that is individuative of *O*. It doesn't matter, at least not in this context, whether we use the term "concept" or "property” when referring to that thing.

[177] Many authorities, e.g. Dummett (1973), believe the principle just stated to be part of the meaning of Frege's celebrated "context principle." I have no doubt that they are correct. It seems to me that Frege's "context principle" consists of three theses:

> (i) For E to refer to O is for ‹...E...› to mean that O is a thing x such that....x...

> (ii) Some expressions, most notably quantifiers, are such that, if one tries to see them as devices of reference or one tries to understand them in any way other than by saying what sentences containing *ipso facto* mean, one will not only be making terrible semantic mistakes, but also terrible logical mistakes that lead to terrible metaphysical mistakes that lead to terrible epistemological mistakes. (See Chapter 1.)

> (iii) Expressions that don't refer cannot possibly be understood except in terms of the contributions they make to the meanings of the complex expressions, and in particular the sentences, in which they occur.

Taken together, (i)-(iii) say that, just as we should see the Earth as revolving around the Sun, and not *vice versa*, so we should see the semantics of sub-sentential expressions as revolving around those of sentences, and not *vice versa*. Put another way, the fundamental law of semantics is that sub-sentential expressions whatever *raison d'être* a given sub-sentential expression has is identical with its ability to contribute to sentence-meaning.

[178] "SS" is short for "Slingshot."

[179] Give or take some irrelevant details, this argument was put forth by Frege in his landmark paper *On sense and reference*.

[180] See his papers "Truth and meaning" and "True to the facts," both of which were published in his (1984) collection of his papers *Essays on Truth and Interpretation.*

[181] Oxford University Press, 2005.

[182] In what follows, I use actual quotation marks where I should, strictly speaking, use ***quasi-quotation marks***. See the Preface, Footnote 8, for a definition of this term.

[183] See Barwise and Etchemendy (1987).

[184] Strawsonians will agree that TF and TF* are equivalent (in the sense that the one cannot be true if the other is false, *not* in the sense they encode equivalent propositions). But they hold that the occurrence in of "this sentence-token" in TF presupposes, and does not affirm, that there exists a unique contextually salient sentence-token.

[185] "RP" stands for "Russell's Paradox."

[186] See Russell (1908).

[187] See Priest (1995).

[188] For reasons to be discussed in Chapter 7 of my book *Analytic Philosophy*, and vaguely adumbrated in Chapter 2 of this book, this means that statements about numbers must either be of Order 3 or of some higher order. Basically, numbers are properties of properties. (The number two is the property had in common by all and only those sets S such that, for some x and y, $x \neq y$ and x and y belong to S and anything z that belongs to S is identical with either x or y. Since the number two is a second-order property, statements *about* the number two are of Type_3.)

[189] Quine (1966) makes this point in the paper entitled "the Ways of Paradox," which is also the title of the book in which that paper occurs.

[190] Short for "alternate interpretation of RP."

[191] What follows is a repetition of the main points concerning context-sensitivity made in Chapter 3. I am doing this to make this chapter as self-contained as possible. In any case, the rest of Section 3.2 may be skipped by those who have absorbed that material.

[192] Parsons (2000) is an example of such a philosopher. Keefe (1997) contains many articles, by various philosophers, advocating the view that there is objective vagueness. In his unpublished (UCLA) dissertation, which was supervised by Kit Fine (himself an advocate of the view that there is objective vagueness), Paul Hovda (2000) argues that there is objective vagueness. (His argument is simply that it's implausible to suppose that the world is articulated into objects all of which have sharp edges. He doesn't consider the argument presented here.)

[193] "OV" stands for "objective vagueness."

[194] “SB” is short for “Smith is bald.”

[195] “ER” is short for “entailment relations.”

[196] Or, more precisely, any given one of the infinitely many possible ways.

[197] Examples of predicates are “tall,” “slow,” and “prime.” Predicates are typically defined as “expressions that pick out properties.” But this definition, though decent enough for some purposes, won’t do as it stands. For there are non-predicates: e.g., “the property of being tall”, that pick out properties. It is not easy to come up with a viable definition of predicate; for any such definition will presuppose the truth of substantive theoretical claims. In any case, here is (what I believe to be) a viable definition of that term. A predicate is an expression *E* such that *E* is elliptical for an open sentence that has the form “*x* is an instance of the property of being *P*” and that, because it has the form, yields an actual sentence when the variable is replaced with either a referring term (e.g., “John”) or a quantifier (e.g., “some penguin”). So “tall” according to this view, is a predicate because it is elliptical for (a grammatical variant of) the expression “*x* has the property of tallness” which, when the variable is replaced with a referring term or quantifier, yields a *bona fide*sentence (e.g., “John has the property of tallness"). The resulting sentence is abbreviated into one that, at the level of orthography, has the form “*x* is *P*.” (So, for example, “John has the property of tallness” is abbreviated into “John is tall.”) This theory—though artificial in some ways---does explain how exactly predicates (e.g., “tall”) differ from the corresponding noun-phrases (“the property of tallness”) and, so far as I know, is otherwise consistent with the data.

[198] "PP" is short for "pure parthood."

[199] The paradox of the heap was first Sorites paradox to come to the attention of philosophers.

[200] I say “relatively” because I don’t wish to rule out that if, given some multiplicity of grains of sand, the distance between any given two of those grains of sand is sufficiently small, that multiplicity could still constitute a heap. If the distance between any two grains is small enough that, as far as the physical behavior of that heap is concerned, those two grains *might as well be contiguous*, then that multiplicity would constitute a heap. That said, to my knowledge, it isn’t physically possible for a multiplicity of non-contiguous grains of sand to have the dynamic properties of a multiplicity of contiguous grains of sand. More importantly, it could be argued that, if a multiplicity of grains of sand *behaved* as though its constituents were contiguous with one another, then they *ipso facto* would be contiguous with another---that the forces binding the grains of sand together, whereby those grains of sand jointly constitute a single object, *make* those grains of sand be contiguous, at least by one legitimate metric.

[201] For a definition of this term, see Chapter 6, Section 4.2.

[202] Wittgenstein (1958) and new-Wittgensteinians, such as John McDowell, hold this. See McDowell (1994, 1998). Donald Davidson has argued for this thesis on several occasions. See his book *Inquiries into Truth and Interpretation.*

[203] The most well-known advocate of this view is Jerry Fodor. See Fodor (1975, 1981, 1987). See also Fodor and Pylyshin (1987) and Pylyshin (1984).

[204] Fodor (1998b) makes similar points (though not quite coincident ones).

[205] Fodor (1998b) gives arguments similar to, if not coincident with, this argument and the next.

[206] Frege (1918) said that, in all likelihood, *all* human beings cannot reason without the help of mental imagery.

[207] There isn't anything it is like to believe that 2 + 2 = 4, but there is something it is like to experience a pain or a pleasure or a tickle. The latter, but not the former, thus have "phenomenology"—they are "phenomenologically pregnant" or, put another way, have "phenomenal content." Anything that there is anything it is like to experience *ipso facto* has phenomenal content (i.e., is phenomenologically pregnant).

[208] Robert Brandom (1994) accepts Wittgenstein's position. And, on the basis of it, he arrives at the conclusion that individuals *never* think—that it is only whole societies that do so. Brandom doesn't consider any of the points made in Section 6.0.

[209] "Quale" (plural, "qualia") is a sensation or, more generally, anything that is felt.

[210] Hume (1740, p. 252).

[211] He does this in Wittgenstein (1958).

[212] "RF" stands for "rule following."

[213] Like this argument of Wittgenstein's, David Hume's argument for his analysis of personal identity assumes that there is nothing more to a person's mind than the images and sensations that populate his consciousness. We'll consider Hume's analysis, as well as his argument for it, in Chapter 16. But Hume's argument doesn't depend as deeply on this assumption as does Wittgenstein's. There is an argument very similar to Hume's, for a position very similar to the one it's supposed to establish, that doesn't involve that doubtful assumption. But nothing even remotely like Wittgenstein's argument is free of that assumption, and there is not, as far as I can tell, any interpretation of Wittgenstein's argument, no matter how liberal, whereby it comes close to making a case for anything at all plausible, let alone accurate.

[214] And, Wittgenstein believes, it's thus a case of the "homunculus fallacy." Searle (1992) regards this argument of Wittgenstein's as probative. In Kuczynski (2007), I thoroughly critique Searle's argument that, in positing the existence of sub-personal thought, one is guilty of what Searle calls the "homunculus fallacy."

[215] Actually, TMA was first put forth—extremely clearly, I might add—by Plato, in his dialogue *The Parmenides.*

[216] Plato himself encouraged this misinterpretation of his own doctrine, since he tried to understand the instantiation-relation in terms of relations (e.g., covering, resemblance, part-hood) that spatiotemporal entities bear to one another. See his dialogue, the *Parmenides.*

[217] Some properties would appear to be instances of themselves (e.g., the property of being a property). But Russell (1908) disagrees. See Chapter 4 of the present work for a discussion of Russell's reasoning.

[218] Readers familiar with Chomsky (1965) will see that I'm using his distinction between performance and competence. The argument being put forth is a highly Chomskyan one.

[219] "WA" stands for "Wittgenstein's argument."

[220] Ayer (1968) makes the same point. Ayer (1968) is, in general, a powerful critique of the private language argument.

[221] We use the expression "the English language" to refer to what are in fact distinct, but connected, languages. The language that Chaucer spoke is an ancestor of the language that we speak. The same is true, though less obviously, of the language that Shakespeare spoke. A language, in the strict sense, is a set of semantic rules.

[222] "DN" is short for "degrees of necessity."

[223] Obviously the *sentences* used to express it aren't necessarily true. It's the content of those sentences that cannot fail to be true. The linguistic vehicles can easily fail to be true; even in our universe, they're only regionally true.

[224] "HIP" is short for "Hesperus identical with Phosphorous."

[225] An explanation of this is given in Chapter 8.

[226] The argument about to be given is found in Bonjour (1998) and also in Dummett (1973).

[227] This is a shame. Quine was a brilliant man. He had a lot to say, and he would have had a lot more to say if he had been a bit more willing to reconsider some of his views. Sadly, it's a rare scholar who is willing to admit that he's wrong.

[228] "DN" stands for "double negation."

[229] See Jerry Fodor's superb paper, "The present status of innateness controversy," in Fodor (1981b).

[230] Quine (1970).

[231] Frege (1956).

[232] In what remains of this chapter, all of the indented expressions are sentences or open-sentences. For that reason, it isn't necessary to enclose an expression in quotation-marks to identify it as an expression that picks out

a sentence, as opposed to a proposition; and I've thus not enclosed the indented expressions in quotation-marks.

[233] In this chapter, there are many places where, technically, I should have used ***quasi-quotation marks*** (see the Preface, Footnote 8, for a definition of this term) but where I instead used ordinary quotation marks or didn't use any quotation marks at all. I didn't want an excess of obscure symbolism to obscure matters of substance.

[234] "LC" stands for "logical constant."

[235] The "‹" and the "›" are quasi-quotation marks. See the Preface, Footnote 8, for a definition of this term.

[236] Wittgenstein (1922).

[237] "AAT" is short for "all analytic truths."

[238] See, for example, Bertrand Russell's 1948 book *Human Knowledge: Its Scope and Limits.*

[239] See Gettier (1963). These very examples are Gettier's.

[240] See Chapter 7.

[241] I am using the word "laws" broadly in this context.

[242] See Lewis (1973, 1984) and Kment (2007).

[243] Arguably, the expression "direct experience" is a pleonasm, like "four-sided square", since, so it could be argued, nothing could conceivably be *in*directly experienced. Supposing this true, the last sentence of the

previous paragraph should be: A theory is an explanation of what is experienced in terms of what is not experienced. Or, alternatively: A theory is an explanation of what is experienced, and therefore known, in terms of what is not experienced and therefore not known *unless* the theory turns out to be correct (and to satisfy the other desiderata that knowledge-conducive beliefs must satisfy), in which case it *is* known.

[244] See Churchland (1988).

[245] Here I am heavily modifying and developing the idea that Ronald Dworkin (1977) made, and put to excellent use, in a totally different context.

[246] The whole concept of defeasibility is alien to empiricism, since it involves the concepts of power, which is one that cannot be understood in empiricist terms.

[247] This argument was put forth by John Stuart Mill in *A System of Logic*.

[248] "WA" is short for "Wittgenstein's argument." WA is a paraphrase, not a quotation.

[249] In his often discussed book *A System of Logic*, written almost 100 years before Wittgenstein wrote the *Philosophical Investigations*, John Stuart Mill provides several masterful illustrations of the spuriousness of exactly the sort of reasoning embodied in WA.

[250] See, for example, Searle (1992).

[251] See Freud (1915) for a telescoped rendition of a similar line of thought.

[252] Russell (1917) puts forth a similar argument. Tyler Burge rejects the position for which I am arguing. See Burge (2007).

[253] “KH” is short for “knowledge how.”

[254] Karl Popper (1902–1994) says this in *The Logic of Scientific Discovery*. Hans Reichenbach (1891–1953) says it in *Experience and Prediction.* Rudolf Carnap (1891–1970) says it is in *The Philosophy of Science.* And Carl Hempel (1905–1997) says it in *The Philosophy of Natural Science* and in *Aspects of Scientific Explanation.*

[255] By “prove” he meant “justify.”

[256] "BL" stands for "blind luck."

[257] Empiricism is rooted in a failure to distinguish knowledge-conducive thought from the things that trigger it. (This is shown in Chapter 13.)

[258] I can only speculate as to why this is so. But it is, I suspect, because philosophers of science tend to be extremely dismissive of any viewpoint that grants the existence of unconscious thought or otherwise makes concessions to psychoanalytic thought.

[259] “HA” stands for “Hume’s argument.”

[260] See Hume *A Treatise Concerning Human Nature* Book 1, Part 4.

[261] “HR” is short for “Hume’s response.”

[262] See Popper (1983).

[263] Short for “not likely.”

[264] At least so far as their various beliefs form a consistent set—we’ll set aside the secondary question of how, given discrepancies among their views, the right ones are to be selected.

[265] Speaking frankly, I’m not 100% that I believe what I’m saying in this passage. I believe that it’s a defensible position. But I *do* believe there to be a connection between an explanation's being good and it’s being likely to be correct. I simply haven’t been able to identify that connection. As a result, the attempt to refute skepticism put forth in this chapter is not adequate. But I have left this chapter in, because I think that components of it are cogent and because, were I to get rid of the non-cogent parts, the merits of the cogent parts would be unclear. (I believe, for example, that my refutation of Hume’s analysis of inductive inference is cogent; and I believe the same to be true of my analysis of scientific explanation. What I don’t believe to be cogent is my attempt to base a refutation of skepticism on that attempt.)

[266] This section is threadbare, to say the least, and needs to be developed, lest the system being put forth here run the risk of collapsing.

[267] See Quine (1953, 1960, 1966, 1981).

[268] Whether this point is consistent with the points I make in Chapter 25 is a delicate issue.

[269] For example, by Ernan McMullen, Professor (emeritus) of Philosophy at Notre Dame. He and I communicated extensively about an earlier draft of the present work. In that dialogue he said that correct explanations cannot possibly be bad, in any sense of that word.

[270] One last point: In his book *The Scientific Image*, Van Fraassen argues that, when putting forth hypotheses, scientists ought to aspire to model the data, not to find the truth. I am afraid that I have not read all of Van

Fraassen's book, or even a large part of it. (But I've read enough of it to know firsthand that I have correctly identified its main thesis.) So I do not know to what extent his reasons for saying this coincide with mine. But I can't imagine that his reasons completely fail to overlap with mine. Not having read his book, I can't speak authoritatively as to his views as to what a scientist's attitude towards the truth ought to be. But, having read parts of it and seen first-hand that Van Fraassen is both talented and honest, my strong suspicion is that he does not advocate any sort of nihilism: he doesn't hold that truth is subjective or non-existent, or anything of the sort. He holds, I would imagine, that good hypotheses are *ceteris paribus* more truth-conducive than bad ones, even though he believes that it is *operationally irrelevant* how accurate a hypothesis is when, in the process of trying to determine the nature of the realities with which that same hypothesis is concerned, one is trying to determine how much credence to give to that hypothesis.

[271] You are "genidentical" with your three year old self. In other words, you and he (or she) are the very same object at different times. The exact nature of genidentity is discussed in Chapter 16.

[272] Locke pointed this out in his *Treatise.* It is not a point that a strict empiricist would make. But empiricists were, and are, slow to realize what their position entails.

[273] Strictly speaking, as Hempel (1953) pointed out, additivity is a relation, not a property. Relative to a given a operation (that of mixing liquids or gases), temperature is not additive. But we could contrive operations relative to which it was additive. Here is a trivial example. Given two objects x and y, whose respective temperatures are m, let us say that x and y have been "plunked" if they are equidistant from some object z such that z was deliberately put in its current position and such that z's temperature is the arithmetical sum of m and n. Obviously plunking is not an operation in terms of which anything of value can be learned about the world, and the same is true of other operations relative to which weight is additive. But it's still worth bearing in mind Hempel's point that it's only in relation to some

operation that a given property can be said to be additive. For there are scientifically important properties that are relative to scientifically some scientifically important operations and non-additive relative to other, equally important operations.

[274] "ET" is short for "Euclidean triangle."

[275] "RC" is short for "Richard Chuck."

[276] This means: *assuming that everything else were the same* or *holding all other factors constant.*

[277] Berkeley makes this point in his (1734) book *The Principles of Human Knowledge.* Wittgenstein (1958) makes some remarks that could be interpreted as attempts to make this point.

[278] His arguments for this are put forth in his book *Three Dialogues between Hylas and Philonous.* Some of those arguments are elucidated in his book *the Principles of Human Knowledge.*

[279] For example, Anthony Brueckner, a contemporary philosopher, believes that skepticism is irrefutable, and he has spent much of his career trying to undermine attempts to refute it. He does this on a case by case basis. In other words, he considers individual attempts to refute skepticism, and tries to undermine them. To my knowledge (which, in this domain, is limited), he has not provided a general argument as to why skepticism is irrefutable. My personal feeling is that we would learn more from such an argument—even a failed one—than we would from the alleged failure of specific attempts to refute skepticism.

[280] This is the argument Berkeley puts forth, though not as speedily, in his *Dialogues between Hylas and Philonous.*

[281] Technically, this is an inaccurate way of putting it because it suggests, falsely, that there are absolute rates of change. But the substance of this point is correct.

[282] He was not the first to do this. Galileo (1564–1642) did it before him, and so did Lucretius (99 B.C.–55 B.C.).

[283] Strictly speaking, shape *per se* isn't a primary property. What is a primary property is some *specific shape.* Similarly, taste *per se* isn't a secondary property. What *is* a secondary property is some specific taste—e.g., the taste of a Snicker's bar. The same thing *mutatis mutandis* holds for the other items on these two lists.

[284] Judging by certain passages in his work, Berkeley himself seems to make this claim.

[285] Actually, Hume would say otherwise; and I'm not entirely sure that he's wrong. (See Chapter 16.) But we can set this aside here.

[286] If I'm thinking *P*, but my opinion is that I'm thinking *Q*, then my opinion as to what I'm thinking is wrong. Of course, my having a given opinion is itself constitutive of what it is that I'm thinking. But any opinion as to what one is thinking at a given time inevitably post-dates that time. If I form an opinion as to what it is that at this exact moment (let *t* be the present moment), my opinion won't come into existence until some time (let *t** be that moment) that just slightly follows *t.* And the opinion that I have at *t** as to what I was thinking at *t* can no more affect what I was thinking at *t* than what people do now can affect what the Romans did 2,000 years ago.

[287] In his *Essay Concerning Human Understanding,* Locke made essentially this point.

[288] If rotten meat had had S*, instead of S, people wouldn't have minded keeping it in their caves (or whatnot), and would have been more

willing to eat it. Food often smells rotten before it tastes rotten. But this (alleged) fact is irrelevant in this context, since we're talking about the epistemological, not the emotional, virtues of odors.

[289] Since you're much smarter than those animals, you can *do* a lot more with the relative paucity of information given to you by your low-res perception than those animals could do with their high-res perceptions. But that has nothing to do with what those perceptions *themselves* are saying.

[290] Short for "question begging."

[291] Because the particles composing very hot things move so fast, those particles do a lot of damage: they damage nerve-endings and other organic tissues. Hence the unpleasantness associated with touching such things.

[292] "BR" is short for "Berkeley's reasoning."

[293] Actually, Berkeley claims—totally implausibly—that people agree with his contention that there are no trans-perceptual objects.

[294] To my knowledge (which is limited, given that I'm not a Berkeley-scholar), Berkeley is never given credit for making these discoveries. I am, so far as I know, the first author to do so. But there is no doubt that he made them. See his work *The Principles of Human Knowledge.*

[295] This is a point that Berkeley himself makes very clearly on several occasions. See his *Dialogues Between Hylas and Philonous*.

[296] In this context, I'm not using the expression 'deep structure' in the Chomskyan sense.

[297] For a sentence to be “perspicuous” is for it to be clear in a logical, as opposed to a psychological, sense. So a perspicuous isn't necessarily one that is easy to understand. Rather, it is one whose syntactic structure reflects that of the fact that, supposing it true, it describes.

[298] Karl Popper had a lot to say about the limits of predictability. What I’m saying here doesn’t always coincide with what he says, but much of it is similar to it. See Popper (1983, 1992).

[299] See Chapter 8 of my (2010) *Morality, Politics, and Law* for a thorough discussion of why even an omnipotent being couldn’t create a square circle or make 1+1=3.

[300] There are trivial exceptions to this. For example, if you know that you’ll be dead in 10 years, then you know that you’ll know nothing. But assuming that your knowledge *grows*, you can’t know how it will grow. And since how it grows affects what you’ll do and therefore affects how the world will be, you can’t predict how the world will be—even if, as Einstein thought, it’s completely deterministic. Thus, the very concept of an infallible self-predictor is an *incoherent* one. In other words, it makes no sense to suppose that a thing should be able to predict its own future states with complete accuracy.

[301] I am indeed aware of the oxymoronic character of expressions such as “infinitely short stretch of time” and (referring to the next sentence) “nil time-interval.” Such pseudo-meaningless expression are abbreviatory devices; they “parse out,” and the post-parse sentence is one that doesn’t refer to the infinitely short intervals of time or other such surds.

[302] Short for “doubly impossible.”

[303] No time would have elapsed unless there had been a change. Time is a relation between events. No changes, no events. No events, no time. The fact that time is to be understood in this way spells doom for

Hume's famous eliminitavist analysis of causation and for a number of doctrines in epistemology and the philosophy of science to which it has given rise. See Chapter 17.

[304] We're setting aside trivial correlations; e.g., "anything that is identical with the headache I'm now feeling is also identical with something I want to get rid of."

[305] Short for "micro-physics."

[306] "AR is short for "Ayer and Ramsey."

[307] Donald Davidson (1970) said that the laws of nature are features, not of the world, but of our descriptions of the world. Davidson's position is false and disingenuous. The laws of nature are not artifacts, and everybody knows that.

[308] This point is central to Clarence Irving Lewis's brilliant (1922) work *Mind and the World Order*.

[309] "AP" is short for average person.

[310] Before we move on, the meaning of the italicized expression in (ii) be clarified. Consider the sentence: (S*) Nothing is a square circle; and if *x* is the total number of cars that some person or other owns, and *y* is the total number of people, then *x*/*y*=2.2.

The logical form of "nothing is a square circle" diverges from its grammatical form. (See Chapters 1 and 3.) Grammatically, it has the same form as "John is a bachelor." But what it says, when stated clearly, is not given by a sentence that has form. For what it says is: *given any object* x, *if* x *is square,* x *is not a circle.* So if *S* is the sentence "given any object *x*, if *x* is a square, then *x* is a not a circle; and the average person owns 2.2 cars" then:

(i*) *S* and *S** say the same thing,

and:

(ii*) *S**'s grammatical form doesn't misrepresent its logical form in the way in which *S*'s grammatical form, *by virtue of the fact that it contains occurrence of "the average man"* misrepresents *its* logical form.

So even though, by virtue of its containing the sentence "nothing is a square circle" *S** *does* misrepresent its own logical form, it doesn't do so in a way in which, by virtue of containing "the average man" *S* misrepresents *its* logical form. Hence the italicized part of (ii).

[311] "RI" is short for "rule concerning the word 'instant.'"

[312] Reichenbach (1928) and Hempel (1952, 1965) clearly and forcefully make these points.

[313] I'm using the past tense to represent all three tenses and will continue to do so.

[314] Russell (1948) put forth this analysis of what an instant is. It may originally be due to Alfred North Whitehead.

[315] One can pick out one of those members obliquely; one may identify it as “the *cause* of such and such” or “the effect of thus and such” or possibly in some other, even more oblique way. But until one of those members has, in some way or other, been picked, that instant has not been identified.

[316] “*S*” stands for “series.”

[317] Oftentimes, a given member of an *I*-class will *itself* token the number (or that fragment of it that it is necessary to state so that, given contextual information, people can identify the rest of the number). By virtue of suddenly saying that it’s 2:46 P.M., Nov. 13, 2009, the clock, supposing it is accurate, is tokening the number (or, in any case, *a* number) that gives the position (relative to other instants) of the instant to which that same event belongs.

[318] Carnap (1966) makes this point. Carnap uses a similar example.

[319] See Mackie (1980).

[320] He did this in the *Metaphysics.*

[321] Short for “Aristotle on causality.”

[322] “But what about the ‘fact’ that 1 + 1 = 2. That isn’t a distribution of mass-energy?” True. But, if the word “fact” is to be of any philosophical use, it has to be disciplined a bit. And in quasi-stipulatively saying that “facts” are distributions of mass-energy, I am deviating from ordinary usage only as much as is necessary to allow that word to have a role in philosophical discourse.

[323] In Chapter 22, functionalism is discussed at length, that being why my discussion of it here isn’t as rich in illustrations and clarifications as

it would otherwise be.

[324] If I have a book that I don't read, but that I routinely use as a weapon, my book *qua* weapon gets plenty of use, but my book *qua* thing that is supposed to be read gets no use. In general, to say that *x qua* phi does such and such, but *x qua* psi does not do such and such is to say (i) that *x* has both phi and psi and (ii) that it is in its capacity as a thing that has phi, and not in its capacity as a thing that has psi, that *x* does such and such. So to say that the *qua* mental has no effects is to say that, although the physiological entities that are presumably identical with mental states have powers, they do not have *psychological* powers; in others, it is not *as* mental entities that they do anything.

[325] See Jackson and Pettit (2004).

[326] To say that a certain interpretation of a term *validates* certain sentences in which it occurs is to say that, given that interpretation, those sentences are correct. An interpretation of a term is an assignment of meaning to it.

[327] "SA" stands for "Smith arson."

[328] Of course, "smaller" doesn't always equal harder to know. It's easier to know people than it is to know galaxies.

[329] Short for "Hume's analysis of causality."

[330] "HC" is short for "Hume on causality." The "1" corresponds to the fact that this is the first of two analyses of causation that Hume gives.

[331] Hume, *A Treatise of Human Nature*, Book 1, Chapter XIV.

[332] See Infeld (1950), Einstein (1952), Reichenbach (1957), Einstein

and Infeld (1967), and Sklar (1974).

[333] See Reichenbach (1957) and Sklar (1974).

[334] See Reichenbach (1957) and Sklar (1974).

[335] “IP” stands for “immediately precede.”

[336] Technical aside: According to Relativity Theory, it makes no sense to say that there is some *one* interval during which *x* and *x** move towards *y* and *y**, respectively. But in this context, we're setting that fact aside. What we are saying is easily adapted to accommodate it. But the adaptations would involve distracting technicalities and therefore aren't worth making.

[337] *"Ceteris paribus"* means "other things being equal" or "other, relevant factors being held constant." Even though *x* and *y* are now closer to each other, it might take light *more* time than it used to travel from *x* to *y*; for *x* and *y* may have been submerged in water or a wall might have been placed in between them. So, given only that *x* is now closer to *y* than it used to be, what follows is that, assuming there to have been no changes of this kind, light takes less time to travel from the one to the other than it used to. In other words, what follows is that *ceteris paribus* this now takes less time.

[338] Replace the word "energy" with the some other term (e.g., "effort,” "time"), if you think that, in using the word "energy,” I am begging the question against Hume.

[339] Short for “no forces.”

[340] Short for “counterfactual analysis of causality.”

[341] Short for “laws of physics.”

[342] See Lewis (1984).

[343] “KN” is short for “Kangaroos with no tails.”

[344] Short for “otherworldly kangaroos.”

[345] “‘LA” is short for “Lewis’s analysis of counterfactuals.”

[346] Short for “standard objection.”

[347] “LC” is short for “Lewis on causality.”

[348] “BU” is short for “balls in the urn.”

[349] “BB” stands for “black ball.”

[350] See Lewis (1986).

[351] “LA” is short for “Lewis’ analysis of induction.” See Lewis (1973, 1984, 1986).

[352] “HA” is short for “Hume’s argument.”

[353] “SB” is short for “stilts beanie”

[354] Goodman (1983).

[355] “GA” is short for “Goodman’s argument.”

[356] Short for objection to the author’s use of Goodman’s

argument.

[357] See Chapter 16.

[358] Hempel (1965).

[359] The “A” stands for “accidental.”

[360] Ayer (1972).

[361] Hempel (1966).

[362] The “N” stands for “nomic.”

[363] The “W” stands for “wide-scope,” the idea being that the nomic operator is given maximally wide-scope, since the whole of “for any *x*, if *x* is metal, then expands” is governed by it.

[364] "SS" and "OS" are short for "scientific statement" and "ordinary statement."

[365] Pierre Duhem (1861–1916) was a physicist and philosopher who Quine wrongly thought to have advocated a similar thesis. In actuality, Duhem advocated a much more circumscribed position.

[366] Hempel (1965).

[367] “NE” is short for “non-explanation.”

[368] “YP” is short for “yellow pencil.”

[369] In this context, “law” and “statement that expresses a law” are used interchangeably.

[370] Because, to put it more generally and also more accurately, a thing’s being in region *R* provides no *nomic basis* for its being yellow. “Nomic” means “having to do with natural law.” If *x* causes *y* to occur, then *x* provides a nomic basis for *y*’s occurrence. But, for reasons that lie outside the scope of this chapter, one state of affairs can provide the nomic basis for another without causing it. Henceforth, the word “causal” will be an abbreviation for “causal *or* nomic,” and the same thing *mutatis mutandis* of “cause,” “causation,” etc.

[371] “VDN” is short for “the viewpoint of DN-advocates.”

[372] “MC” stands for “the minimization of causal anomalies.”

[373] “ED” stands for “elimination of discontinuities.”

[374] “OH” is short for “objector’s hypothesis.”

[375] In this context, *anything* described by a true sentence counts as a “state of affairs.” So facts are “states of affairs” as we’re using this expression, and so are occurrences.

[376] "IP" is short for "intrinsic probability."

[377] To be precise, they are, ultimately, known to us through sense-perceptions and therefore through non-propositional information. But that non-propositional information must be propositionalized in order to engage the system of propositions constitutive of a science. So, given that, in this context, we’re analyzing the concept of scientific explanation, and given that, as far as science is concerned, events must be propositionalized if they are to be amenable to scientific explanation, we can, speaking less than entirely

accurately but without diminishing the cogency of what we are saying, speak *as though* events were given to us through propositions.

[378] Short for “plausible objection.”

[379] *x* and *y* are *genidentical* if *x* is an older version of *y* or *vice versa.* You are genidentical with your three year old self. Since you have properties that your three year old self did not have, we can't say that you are *identical* with your three year old self either rejecting Leibniz's Law or accepting some non-trivial semantic and metaphysical contentions. The word "genidentical" enables us to avoid prejudging the issues that we'd have to have settled before we justifiably accept any of these contentions. In Chapter 20, it is made clear that the correct path is to modify Leibniz's Law---to hold that, although it's true for "synchronic" identity, it is not true for "diachronic" identity.

[380] “CA” is short for “classical answer”

[381] To my knowledge, Bernard Wiggins was one of the first, if not the first, to make this distinction. He did so in his (1968) article “The statue and the clay,” which was published in *Mind.* The example I’m using to illustrate this distinction is Wiggins’ own.

[382] Many philosophers of science reserve the term “law” for the sorts of exceptionless regularities not found outside of physics and some branches of chemistry. Relative to that usage of the term “law,” there are no laws of botany. But partly because, as we’ve seen, laws aren’t regularities, and partly because it’s convenient to do so, I’m going to talk about “laws” of botany and psychology and other disciplines in which there are no such regularities, even though, in so doing, I may not be speaking entirely accurately.

[383] “CW” is short for “conventional wisdom”

[384] “HP” is short for “hypothetical penguins.”

[385] We are setting aside as irrelevant the fact that, because of interference-effects, it wouldn't be possible, even after your limitations as a human being are set aside, for you to know, at any given time, how any two micro-particles composing a given object at *t* were interrelated.

[386] I should point out (i) that the claims just made are empirical and therefore could be false, (ii) that I am far from being an expert on brain-science, and (iii) that Jaegwon Kim (1993), whose judgment I have found to be accurate in most cases, holds that what I just said is false (Kim is a philosopher, not a physiologist, and rejects what I said on *a priori* grounds, not on empirical grounds, which reveals the interesting fact---which we don't have time to pursue, unfortunately---that, in Kim's view, it is *not* an empirical question whether the psychological is multiply realizable). These disclaimers are not formal. I invite the reader to look into (i) and see if I was right to judge that contemporary research bears it out. (I made (i) on the basis largely of the testimony of medical professors. I have no first-hand knowledge of this area.)

[387] See Chapter 9, Section 9.0 for definitions of these terms and explanations of the corresponding concepts.

[388] One's preconscious comprises everything in one's mind of which one readily become conscious. So my knowledge that 2 + 2 = 4 belongs to my preconscious: I am not always conscious of it; but I am always capable of instantly becoming conscious of it. One's preconscious mind does not include the repressed; nor, obviously, does it comprise the sub-*personal* processes, such as those involved in language-comprehension and sense-perception.

To my knowledge, Freud (1915) was the first author to use the term "preconscious" (or its German counterpart, rather) in this way. At any given moment, most of what is in a given person's preconscious is unconscious; i.e., he is not, at that moment, conscious of it. Since, given any one of those contents, that person could readily become aware of it, it follows that its being unconscious has to do with the fact that his attention is elsewhere: the spot-light of his consciousness happens not to be shining on that bit of mental

topography, even though there is nothing preventing it from shining on it. Freud described the repressed as "dynamically" unconscious. One's failure to be conscious of some repressed item is not the result of the fact the search-light of his consciousness is focused on the wrong part of his mental terrain. It is the result of the fact that there are forces that actively keep it away from any place on which that search-light can shine.

Chomsky described sub-personal mentation as "deeply unconscious." What is repressed can be made conscious. But what is sub-personal cannot be made conscious. Since, there, the sub-personal is *more* removed from consciousness than the repressed, it makes sense to describe it as "deeply unconscious."

[389] In his book *Essays on Actions and Events.*

[390] “DA” is short for Davidson’s argument.

[391] Not all propositions are explanatorily robust. As we’ll see in a moment, some propositions are such that one couldn’t even *conceive* of scenarios in which they were explanation conducive.

[392] Each corresponds to infinitely many concepts, actually.

[393] Strictly speaking, I don't think that this is the right way to think about concepts. See the end of Chapter 2.

In this context, in saying that Smith “directly” knows that (for example) there is a plant in the basement, I am saying that he knows it *relatively* directly. The sense in which he knows it directly is comparable to the sense in which I know directly that my car is in the driveway. In saying that Smith knows it "directly," I am not saying that Smith did not at any level have to make any inferences to acquire that knowledge (which, given what we saw in Chapter 9, is unlikely if not impossible); nor am I saying that any such inferences as Smith did make in acquiring that knowledge were

themselves members of *S#*.

Given that, to those who haven't read Chapter, the preceding words may be very obscure, it may be appropriate to say a few words as to what they mean and why they're justified. To know that the thing in my basement is a plant, I may have to know that I am not prone to hallucinations of plants, that my basement isn't equipped with a device that creates plant-holograms, etc. So it may be that my knowledge that I am seeing a plant is indirect in the sense that, *given* what my visual perception is telling me, *along* with my background knowledge as to the lighting conditions in the basement (etc.), I am inferring that there is a plant. But the knowledge involved in that inference (in addition to being made at a sub-personal level and, therefore, not being an inference that *I* make) doesn't belong to *S#*. So *relative* to *S#*, my knowledge that there is a plant in my basement is a datum.

The concept of datum-hood is relative in many senses. The context warrants yet one more repetition of this point. One cannot literally *see* that something is a *bona fide*plant as opposed to a non-plant whose surfaces reflect light in exactly the same way as those of a *bona fide*plant.)

[395] I don't believe that class to be empty. I believe that class to include a sentient being's knowledge of certain facets of his consciousness and a sapient being's knowledge of the truth-conduciveness of certain rules of inference. For reasons explained in Chapter 9, I believe that, unless we had such intuitions, we couldn't know anything. So there are, I believe, *some* truly "hard" data. But practically nothing that anyone, scientist or layperson, describes as a datum falls into that category. Any such thing is, at most, only a datum in the multiply relative sense earlier described.

[396] Advocates include: Plato, Descartes, Leibniz, Karl Popper, and Saul Kripke.

[397] Advocates include: Aristotle, Thomas Hobbes, J.J. Smart, Hilary Putnam, and Jerry Fodor.

[398] Spinoza advocated this view.

[399] Quine (1960) advocates this view.

[400] In *The Principles of Human Knowledge*, George Berkeley advocates this view. Other advocates include Bertrand Russell (*The Analysis of Matter*), and William James (*Essays in Radical Empiricism*).

[401] In the *Monadology*, Leibniz gives a similar argument, and so does Gottlob Frege (1918).

[402] Frank Jackson forcefully makes this point in several articles. See Jackson (1998).

[403] In his paper, "Mental Events," Donald Davidson argues for this position, or one close to it. We stated and evaluated Davidson's argument in Chapter 21, Section 5.4.

[404] Hume and Hempel hold this. See Chapter 21.

[405] The following argument is due to Quine (p. 153 of *From a Logical Point of View*).

[406] Hempel (1952) and Pap (1958) made this point. Kripke (1972) and Putnam (1975) put it to good use.

[407] J.J. Smart (1959) gives this argument. So does Fodor (1968).

[408] C.D. Broad makes a similar point in *The Mind and Its Place in Nature*.

[409] My own view is that this reasoning is solid. I say this despite my

unshakable, but—so far as I know now— very hard to defend suspicion that dualism cannot be completely wrong.

[410] In Chapter 24, it is said what exactly it means to identify objects with causal sequences, and it is also said why exactly objects must be identified with such sequences.

[411] Fodor (1968).

[412] Richard Rorty (1979) makes this point.

[413] Searle (1992) clearly states that the word "subjective" is ambiguous in the way just described.

[414] See Watson (1924) and Carnap (1934).

[415] “RS” is short for “Ramsey sentence.” RS is given by David Lewis in his (1983) paper “Psychophysical and Theoretical Identifications.” See also the Chapter in Carnap (1966) titled “The Ramsey Sentence.”

[416] “ARS” is short for “anything representationally symbolic.”

[417] This must be qualified. It’s usually not a good idea to look *primarily* to legal principles for answers to philosophical questions. But having knowledge of legal principles can help one fine-tune the answers one has already come up with. For example, many of the laws concerning the admissibility into legal proceedings of certain kinds of evidence do embody real insight into the structure of knowledge.

[418] In his (1949) work *the Philosophical Investigations,* Ludwig Wittgenstein said that all philosophical problems are to be solved by looking at how words are used. (In Chapter 1, it is argued that this view is untenable.)

This work of Wittgenstein's initiated a philosophical methodology known as "ordinary language philosophy." (Among the proponents of ordinary language philosophy are J.L. Austin (1911–1960), Norman Malcolm (1911–1990), and Peter Strawson (1919–2006). Strawson was a truly great philosopher, but he was stunted by the influence that Wittgenstein had on him. I have to admit that Austin *did* do some very good work that was influenced by Wittgenstein. I do not say this because I am reconsidering my negative take on Wittgenstein. I am simply registering a fact.) Ordinary language philosophers—including the ones just mentioned—put a lot of stock in arguments such as the one just given. Their position would be that, since we can refer to corpses by the names of their former occupants, those corpses *are* those occupants. Although there is much to be learned from how to use words, it is quite absurd, in my judgment, to hold that philosophical problems are categorically to be solved by studying linguistic behavior.

[419] In his (1973) book *Problems of the Self*, Bernard Williams (1929–2003) puts forth a similar argument.

[420] "TF" stands for "things are fundamental."

[421] "PB" stands for "property instances are basic."

[422] One of Hume's great contributions to philosophy was to expose our tendency to do this, along with the errors to which it gives rise.

[423] Quine (1960) advocates this view.

[424] In his (1956) book *Elements of Symbolic Logic*, Rudolf Carnap uses the term "time-slice" (or its German equivalent, rather) in this way. To my limited knowledge, he was the first to do so.

[425] See Parfit (1984).

[426] See Chapter 13.

[427] See Chapter 17 for a discussion of how, exactly, Hume's views on these matters were conservative.

[428] Hume (1740: 252).

[429] Hume (1740: 252).

[430] Defeasible" means "capable of being overridden or proven false." Etymologically, it means "capable of being defeated."

[431] Also, any given spatiotemporal thing is moving relative to something.

[432] Obviously (i) isn't *formally* false. It isn't like "snow is white and it is not the case that snow is white." Nor is (i) tautologously false. It isn't like "there are not three feet in a yard." (i) is *analytically* false. It embodies an incoherent assumption about the nature of persistence in time. This assumption is discussed in Chapter 15.

[433] To use Fodor's apt expression. See his paper "The Present Status of the Innateness Controversy," in Fodor (1981).

[434] Kant makes a similar point at the very beginning of the *Critique of Pure Reason.*

[435] See Chapter 22.

[436] Jean Paul Sartre (1905–1980) and Maurice Merleau-Ponty (1908–1961) are dualists who believe in free will and hold that the physical world is deterministic, and Peter van Inwagen (1942–) is Peter van Inwagen

(1942–) materialist who believes in free will and hold that the physical world is indeterministic.

[437] See Freud (1949).

[438] The straightforward version ("you are free if you can do what you want to do, and unfree so far as you cannot") was advocated by John Locke (1632–1704) and G.E. Moore (1873–1958). The more sophisticated version is advocated by Immanuel Kant (1724–1804) (in the *Critique of Practical Reason*), Sigmund Freud (1856–1941) (in *An Outline of Psychoanalysis* and *the Question of Lay Analysis*) and Harry Frankfurt (1929–) (in *The Importance of What We Care About*).

[439] Moore (1911).

[440] In philosophy, split infinitives are hard to avoid, since the consequences of avoiding splitting them are often sentences that don't mean what one wants to say. There's a difference between saying "Smith chose to *not* strike Harry" and "Smith did not choose to strike Harry." The latter says that Smith failed to make a choice; the former says that he did make a choice. Of course, "Smith chose *not* to strike Harry" does avoid the split infinitive. But having the "to" adds oft-needed emphasis.

[441] See his *Essay Concerning Human Understanding*, Chapter XXI.

[442] "CR" is short for "Chisholm's response."

[443] See Frankfurt (1988).

[444] "SM" stands for "Smith microchip."

[445] This is thoroughly discussed in Chapter 17.

[446] "PAP" stands for "principle of alternate possibilities." Frankfurt coined this term.

[447] More precisely: if JFK hadn't been assassinated, but *otherwise* things were the same *to the extent that, given that JFK wasn't assassinated, they could be the same*, a consequence being that JFK was healthy and popular, etc., *then* he would have been re-elected.

[448] There are exceptions to this. For example, counter*logicals* aren't causal claims. An example of a counterlogical would be:

(*) If triangles had four sides, they'd have an even number of sides.

(*) obviously isn't a causal claim. (*)'s meaning parallels (JFK)'s. Neither says anything about other worlds. Each merely registers some dependency that holds in this world.

[449] "FA" is short for "free act."

[450] See Frankfurt (1988, Chapter 2).

[451] Frankfurt's analysis of freedom corresponds very closely to the conception of freedom underlying the work of Freud and other thinkers in the psychoanalytic tradition. But Frankfurt was, to my knowledge, the first person to articulate this conception of freedom, and he was the first to defend it in a systematic and rigorous manner.

[452] For an act to be rational, it is to satisfy two conditions:

(i) The agent must have good reason to believe the immediate objective of that act to be one that serves his interests

and

(ii) The agent must have good reasons to believe that, given the way in which he is attempting to attain his immediate objective, he is indeed likely to attain it; he must believe his methods to be efficacious.

An *intelligent* act---or, to put it in a way that will probably lead to less confusion, an act that *shows* intelligence---need not satisfy the first condition. An intelligent act *does* have to satisfy the second condition. But must also satisfy another condition, namely: (iii) An unusually high degree of acumen on the agent's part was involved in his being justified in believing that this methods are likely be efficacious.

[453] "FD" can be thought of as short for "freedom to do what one wants."

[454] Addiction doesn't obliterate a person's personality; it only temporarily obscures some of the outlines of the psychological architecture of

a person’s identity. The addiction may lead to some character remodeling. But, in most cases, the person is still there, even at the height of the addiction.

[455] Freud, Stekel, and other early psychoanalysts referred to OCD as “obsessional neurosis.”

[456] .See Davies (2001, 2001b).

[457] A colleague of mine (who wishes to remain anonymous) told me that sickle cell anemia is adaptive in some respects. My claim that physical ailments aren’t adaptive must obviously be qualified. Many ailments involve over-active immune systems and can be beneficial (in certain respects, at certain junctures).

[458] Some philosophers deny this and so do and many theoretical as well as clinical psychologists. In their view, the very idea that a person should rationalize or otherwise deceive himself is incoherent and is thus false *a priori*. In Chapter 27, we will see why such thinkers are wrong.

[459] See *The Ego and the Id* (1923) and *Inhibitions, Symptoms, and Anxiety* (1926).

[460] He’s right about this, in my judgment; but my reasons for agreeing with Mill on this are not---at least not according to the traditional interpretation of the relevant works of Mill’s---in agreement with Mill’s own reasons.

[461] Many people who focus on martial arts (as I used to do) are, I suspect, in this category. They're substituting the development of hand-to-hand combat-skills—which, since pre-historic times, have had only symbolic value, being of no use in increasing one's actual standing in the world, except in the rare case where one becomes a professional martial artist.

[462] The only exception to this is ethical egoism, which we’ll discuss

later.

[463] In any case, there is nothing under whose jurisdiction they fall *by virtue of* their being occupants of *LM*.

[464] The contents of this section are not definitive. Here I am investigating ways of analyzing concepts (self-deception and repression) that I know (for reasons that lie outside the scope of this chapter) to have integrity, but whose structures have not thus far been made explicit and upon whose integrity it has thus far been possible for the many opponents of psychoanalytic thought to cast unwarranted aspersions.

[465] "OR" stands for "opponents of repression."

[466] I am using this term in the sense in which Freud used it, though my analysis as to what precisely is involved in it is not, so far as I know, to be found in Freud's work *per se*. My analysis is an attempt to make explicit the philosophical assumptions underlying Freud's work and that, when made explicit, prove to be quite coherent.

[467] In the last sentence, I was, for expository reasons, setting aside the odd (or, actually, not so odd) theoretician whose relation to his work becomes entangled with personal animosities towards colleagues or otherwise becomes emotionalized. I know that, out of dislike for some specific individual who was associated with a certain theory, I was reluctant to accept it. In each such case that I can recall, I did eventually accept the theory in question—but only because what I stood to gain outweighed the humiliation of having to concede that this antagonist of mine was right.

[468] Why the parenthetical hedge? Because I believe that, in many cases, one unconsciously makes the inferences that, through such alterations of one's life, one is trying to avoid making. But since it would, arguably, be question-begging of me to assume that here, the occurrence of "consciously" occurs in parentheses.

[469] See the previous note.

[470] Quine holds that, given any body of data, there are infinitely many *equally* reasonable ways to model it. I argue in Chapter 20 that Quine's position is either false or trivial, depending how it's interpreted. But the body of data involved in one's making a delicate judgment call (such as is involved in judging somebody to be a psychopath or, more generally, in judging somebody to fall under this as opposed to that psychiatric rubric), the body of data involved in one's making an inference to the best explanation is such that it can reasonably be modeled in more than one way.

[471] "SA" is short for "Sartre's argument."

[472] Given the points just made, this means that $S_1 \ldots S_n$ satisfies two conditions. First, for each *i*, any two members of S_i are capable of directly pooling their contents. (In other words, they can pool their contents in the way in which the contents of my perception of Jerry's grinning face can be pooled with the contents of my beliefs about Jerry, meaning that no overt behaviors or other such intermediaries would be involved in such a pooling.) Second, for any *i* (between 1 and *n*, inclusive) and any *j* (same qualification), the contents of any member of S_i can be pooled, in the (relatively) direct way described earlier with the contents of any member of S_j.

[473] Nor was Saddam Hussein. Nor was Hitler. These people were thoroughly evil. But they were not psychopaths. Strange as it may sound, they were *too* evil to be psychopaths. Psychopaths don't have the psychological integrity to be evil in the same way as, or to the same extent as, the likes of Saddam, Hitler, Stalin, Mao, and Pol Pot.

[474] See Cleckley (1964), Chapter 58.

[475] Cf. Cleckley (1964), p. 409: "Without suffering or enjoying in significant degree the integrated emotional consequences of experience, the

psychopath will not learn from it to modify and direct his activities as other men whom we call sane modify and direct theirs. He will lack the real driving impulses which sustain and impel others towards their widely differing but at least subjectively important goals."

[476] Cf. Cleckly (1964), p. 406: "Logical thought processes may be seen in perfect operation…All judgments of value and emotional appraisals are sane and appropriate when the psychopath is tested in verbal examinations. Only very slowly and by a complex of multitudinous small impressions does the conviction come upon us that, despite these intact rational processes, these normal emotional affirmations, and their consistent application in all directions, we are dealing here not with a complete man at all but with something that suggests a subtly constructed reflex machine which can mimic the human personality perfectly. This smoothly operating psychic apparatus reproduces consistently not only specimens of good human reasoning but also appropriate simulations of normal human emotion in response to nearly all the varied stimuli of life."

[477] Which Cleckley, in the eighth edition of *The Mask of Sanity*, described as the best fictitious portrait ever of a psychopath. Aston's book is brimming with intelligence and is, in general, quite superb. But Cleckley's endorsement of this work must be taken with a hefty grain of salt. Certain passages of *The Incredible Henry Carewe* make it clear that Aston had read *The Mask of Sanity* and had tailored her novel to fit Cleckley's work. So even though Aston's novel *would* have corroborated Cleckley's views *if* it had been based on her own, Cleckley-independent observations, it fails, for the reasons just given, to corroborate those views. But this doesn't really matter, since corroboration for Cleckley's views is not only not hard to come by, but hard to escape, given the pervasiveness of the phenomenon he is so aptly describing.

[478] Lindner (1944, pp. 5-6).

[479] Cleckley (1964), p. 199.

[480] ‘SR’ stands for ‘Smith’s rationalization.’

[481] The contents of this section are similar to those of Chapter VII of Freud’s (1900) book *The Interpretation of Dreams.*

[482] Freud makes this point in Chapter 3 of his (1930) book *New Introductory Lectures on Pyschoanalysis.*

> [483] What do I mean by the term “sub-agential”? Not everything that happens *in* one’s mind is something that one *does*. It isn’t *my* doing that is responsible for my visual perceptions. Those perceptions just happen. All I have to do is make sure that nothing bad happens to my eyes. So the construction of visual perceptions is sub-agential. *I* am an agent. *I* do not construct my perceptions. So the part of my mind that constructs perceptions is sub-agential. By contrast, the decision that I made to spend this morning writing this chapter, instead of getting an allergy shot, was agential. *I* made that decision. It wasn’t made for me, as it were, by some part of my mind whose behavior falls outside the reach of my choice-making faculty.

[484] If I’ve interpreted him correctly, Nietzsche said what I’m about to say. See his book *The Will to Power.*

[485] I say this on the basis of what many a Stalin-biographer (Conquest, Tucker, DeJonge, Medvedev, Volkogonov, Radzinsky, Kruschev (I’m counting his memoirs as a quasi-biography of Stalin), Bullock, Deutscher, Ulam, and Trotsky, among others) has written. Some of those biographers (e.g., Trotsky and Khrushchev) personally knew the man. (Unfortunately, Richard Pipes, the great cold warrior, never wrote a biography of Stalin.) Ultimately, I found Medvedev’s to be the best overall; but I found Radzinsky’s to be the most insightful. Isaac Deutscher’s biography was the best from both a literary and a philosophical point of view---but, from a historiographical point of view, it wasn’t remotely comparable

to Medvedev's, and from a psychoanalytic point of view, it wasn't quite in the same league as Radzinsky's. (The worst was Ulam's biography. Which is ironic, given that he was considered the leading expert on Stalin for many years.) Also, the biographies of Mao that I've read provide strong, indirect support for my contention that Stalin knew who and what he was.

[486] See the footnote before the previous one.

[487] Actually, Davidson says that intentions are "desire-belief" pairs such that the desire that belongs to that pair is one's strongest contextually relevant desire. Setting aside the confusions embodied in Davidson's contention that intentions are desires, Davidson's belief that intentions are belief-desire pairs makes little sense. My *desire* to drink the ice water in front of me *itself* incorporates my belief that there is ice-water in front of me. There are often cases where one has a desire (e.g., a desire to drink ice-water) but doesn't have any beliefs as to how to gratify it. But once one does acquire such a belief, it is itself incorporated into the desire. When I discover that there is ice-water within my grasp, and that, the circumstances being what they are, I won't get in any sort of trouble for drinking it, what was previously just a desire for ice-water on my part becomes a desire to drink the ice-water *in front of me*. So the belief (*there is ice-water in front of me*) is absorbed into the desire. Thus, supposing for argument's sake that Davidson's theory is at least an approximation to the truth, intentions aren't belief-desire pairs: they're just desires, albeit ones that like many desires, if not all of them, incorporate beliefs.

[488] In any case, I consider the points about to be made a development of Bratman's points, much as I consider my analysis of freedom a development of Frankfurt's analysis. Speaking personally, I accepted Davidson's analysis before I read Bratman's book, just as I accepted Locke's analysis of freedom before I read Frankfurt's work.

[489] Actually, even this way of putting it suggests that Davidson's analysis is less defective than it really is. Davidson's analysis doesn't embody any awareness of the distinction between minds and selves. So, since

the word “us” is a near-synonym of “our selves," it’s misleading to say that, according to Davidson’s analysis, our desires use *us* to express themselves, it being preferable to say that, in order to express themselves, desires use the information contained in the minds that, when they become selves, become us.

[490] It isn’t *necessarily* done shoddily. One damages oneself if one doesn’t gratify intense instinctual urges. (What follows is similar to, if not identical with, a point made by Freud (1923) and Freud (1926.) And---speaking from a medical, as opposed to an ethical, point of view---if an opportunity arises in which one can gratify them without risking incarceration or other forms of damage to oneself, then one *should* gratify them. Those impulses will find a way to discharge themselves. If you thwart them, the pressure that they continue to exert will warp your decision-making process; they will do so by influencing the decisions that you make in such a way that, through the ensuing actions, they be expressed. But, oftentimes, those decisions should *not* be vehicles for impulses: those decisions may concern professional matters to which those undischarged impulses are irrelevant. And, should those decisions come to be vehicles for one’s aggressive or sexual or otherwise primal urges, they will not be good ones.

[491] See Bratman’s brilliant discussion of “bootstrapping” in Chapter 2 of *The Face of Intention.*

[492] In any case, it is the problem that, if I’ve interpreted Bratman correctly, he believes to be the problem with Davidson’s analysis. I cannot be confident that I’ve interpreted Bratman correctly, since his discussion of key points tends to be cautious and highly qualified and therefore obscure. I can say with certainty, first, that, were it not for my having read Bratman’s book, my eyes wouldn’t have been opened to the problems with Davidson’s analysis, which, as I said earlier, I used to find plausible. And I can also say with certainty that much of what I’m saying, much though it owes to Bratman, is not to be found in his work.

[493] For a discussion of ethical egoism, see my (2010) book *Morality,*

Politics, and Law.

[494] People *sometimes* believe that they're frauds. But, in the rare cases when they come to that conclusion, they're slow to do so, and the conclusion reached is not as black as it ought to be. For everyone, including self-admitted frauds, always underestimate the degree to which they are frauds. (To quote La Rouchefoucauld: "People are never aware of the scope of their self-deceptions.") But their confessions to themselves not only *under*represent, but also *mis*represent, what they're admitting about themselves. For, when a person admits to being a fraud in one respect, it is to conceal from himself the fact that he's a fraud in some other, more malignant respect. Most admissions of failure and fraudulence are token.

[495] "Alethic" means "relating to truth."

[496] A complex statement is a truth-function of the statements composing it if its truth-value is a function of their truth-values. The sentence "snow is white and grass is green" is a truth function of the sentences composing it, since it is true if they. "It is not the case that snow is white" is a truth-function of "snow is white," since the former is true if the latter is false. But "snow is white because grass is green" is not a truth-function of its constituent sentences: given only that each of "snow is white" and "grass is green" is true, it's indeterminate whether it is *because* grass is green that snow is white. Thus, whereas "and" and "not" (or "it is not the case that") are "truth-functional operators," "because" is a non-truth-functional operator.

[497] I have known for a while that somebody or other wrote an article with this title; but I didn't know who wrote it, or where or when it was published. I am indebted to my copy-editor for supplying me with this information.

[498] See Ray Monk's (1990) biography of Wittgenstein, *The Duty of Genius.* Monk idolizes Wittgenstein. But, as far as I can tell, he does not white-wash the facts in Wittgenstein's favor. He quotes Russell as saying, during the first few months of his relationship with Wittgenstein, that the

latter was a “fool”, with whom it was a “waste of time” to converse. He also then points out that, for some reason that Monk does not identify, Russell underwent a complete change of heart.

[499] See Dummett’s paper “Can analytic philosophy be systematic, and ought it to be?” published in his (1978) collection of essays *Truth and Other Enigmas.*

[500]. In his (1976) paper “Can there be vague objects?” Gareth Evans says that, according to a viewpoint that he rejects, “rather than being a deficiency in our mode of describing the world, [vagueness] would be a necessary feature of any true description of it.” For reasons of professional ethics, I wish to bring the reader’s attention to the (unintentional, at least at the level of consciousness) similarity between my wording and Evans’.

[501]. Contrary to what Quine alleged, Pierre Duhem (1861–1916) did not put forth the so-called “Duhem-Quine” thesis.

[502]. Though I believe there to be a propositional analogue of it. See Chapter 3.

[503]. There isn’t really such a thing as *the* logical form of a sentence. What it is that counts as a clarification of a sentence’s meaning depends in part on what one’s starting-assumptions are. For example, if one takes the concepts *point, plane,* and so on, as primitive, the sentence “x is the class of points equidistant from a given point on a plane” might give the logical form of “x is a circle.” If one doesn’t take those concepts as primitive, it might abjectly fail to do so.

Also, given some language L that is expressively more powerful than English, the L-sentence that would give the logical form of “snow is white” would not be a direct translation of the English sentence that would do so. So it could be sentence that the logical form of “snow is white” *relative to L* isn’t identical with the logical form of “snow is white” relative to English. (That said, I’m not entirely convinced that logical form is relative in this way.)

[504]. For example, Peter Strawson in his otherwise brilliant (1950) paper “On referring.”

[505]. The sentence “this is radically absurd” occurs in Russell’s (1903) book *Principles of Mathematics* (PM). (It occurs in the chapter on Newton’s laws of motion.) I rather fancy this sentence, and thus use it from time to time. Ironically, the statement that Russell describes as “radically absurd” in PM is, as I recall, quite a defensible one.

[506]. “SS” is short for “Slingshot.”

[507]. Advocates of this argument say that the true sentences refer to *The True*, and that false sentences refer to *The False*. They don’t say what these terms mean. In an effort to free their argument of needless obscurities and indefensible assumptions, I’ve taken the liberty of assuming that they refer to the properties of being true and false, respectively.

[508]. Give or take some irrelevant details, this argument was put forth by Frege in his landmark paper *On sense and reference*.

[509]. See his paper “True to the facts,” in the (1984) collection of his papers *Essays on Truth and Interpretation.*

[510]. Oxford University Press, 2005.

[511]. In what follows, I use actual quotation marks where I should, stricty speaking, use quasi-quoatatoin marks.

[512]. To my knowledge, this term was coined by John Searle. It is the title of his excellent (1969) book *Speech acts.*

[513]. Examples of universal quantifiers are “for all x,” “given any x,” “any object,” “all objects,” and so on. (In general, a universal quantifier is an expression E such that, in virtue of having the form ‹...E...›, a sentence expresses a proposition to the effect that all members of some one class are members of some other class. Thus, “for all x, if x is a dog, then x is a mammal” is to the effect that every member of the class of dogs is a member of the class of mammals.

[514]. I owe much of what I know about virtue theory to conversations with Anthony Ellis.

[H1]{au: if this is a quote than the variables would not be italicized here unless they were in the original, so they have not been corrected WHAT FOLLOWS IS NOT A QUOTE; IT'S A SUMMARY, CONSISTING ENTIRELY OF MY OWN

WORDS, OF A LONG TRACT OF TEXT. WHEN I QUOTE SOMEONE, I MAKE IT CLEAR THAT I AM DOING SO. I NEVER EVEN USE PHRASES SIMILAR TO THOSE THAT I'VE READ WITHOUT IDENTIFYING THE SOURCE OF THE PHRASE RESEMBLING THE ONE I'VE USED. BEST, JM}

[H2]{au: APA generally calls for a.m. and p.m. to be in lower case, however, to save number of corrections required I am correcting to small caps throughout text for consistency OK WITH ME--JM}
[H3]{au: ok? THIS WAS NOT A TYPO. I MEANT TO SAY, AND DID SAY, THAT THE PROPERTY OF BEING A ROCK *DOES* EXIST.}
[H4]{au: isolatable? OK---I WAS NEVER CLEAR ABOUT WHETHER IT WAS 'ISOLABLE' OR 'ISOLATABLE.' I THOUGHT 'ISOLABLE' WAS THE WORD (OR *A* WORD, ANYWAY), BUT I'M NOT 100% ABOUT THIS. YOUR CALL.}
[H5]{au: is this a quote? NOT A QUOTE—A PARAPHRASE}
[H6]{au: is this a quote? NOPE—NOT A QUOTE. A PARAPHRASE. THESE GUYS WERE REALLY LONG WINDED---I'M TRYING TO SPARE MY READERS THE PAIN THEIR PROSE INFLICTED ON ME.}
[H7]**{au: changed TOC, ok? OK--JMK}**
[H8]{au: is this a quote? NO---IT IS NOT A QUOTE, JMK}
[H9]{au: I thought that roulette had one ball, not two dice? or alternatively can be played using dice instead of the wheel to designate the numbers, but not both dice and a wheel? YOU'RE PROBABLY RIGHT---I DON'T KNOW ANYTHING ABOUT GAMBLING. I DON'T THINK INACCURACIES SUCH AS THESE MATTER. BUT DO YOU WHAT FEEL IS NECESSARY.}
[H10]{au: switched to ball instead of dice here?} YOUR CALL, BEST JM
[H11]{au: is this a quote?---NO, NOT A QUOTE}

[H12]{au: are these quotes? NO, THEY ARE NOT QUOTES, BEST JMK}

[H13]**{au: ok? YES, jmk**
[H14]{au: or remove other quotes? REMOVE QUOTES, THANKS---TO HARD TO MAKE IT UNIFORM, ESPECIALLY NOW THAT I'M WAY BEHIND SCHEDULE. }

[H15]{au: are the sentences in quotation marks quotes? NOT NECESSARILY; I LEAVE THIS TO YOUR DISCRETION.}
[H16]**{au: no hyphen in later areas, changed here, ok? Or add hyphen throughout? I LEAVE THIS TO YOUR DISCRETION, TIME RUNNING OUT—BEST, JMK}**